DOI: https://doi.org/10.17816/DD71031

Рекомендации по составлению отчётов о диагностических исследованиях (STARD 2015): разъяснения и уточнения

J.F. Cohen[1, 2], D.A. Korevaar[1], D.G. Altman[3], D.E. Bruns[4], C.A. Gatsonis[5], L. Hooft[6], L. Irwig[7], D. Levine[8, 9], J.B. Reitsma[6], H.C.W. de Vet[10], P.M.M. Bossuyt[1]

[1] University of Amsterdam, Амстердам, Нидерланды
[2] Paris Descartes University, Париж, Франция
[3] University of Oxford, Оксфорд, Великобритания
[4] University of Virginia School of Medicine, Шарлотсвилл, Вирджиния, США
[5] Brown University School of Public Health, Провиденс, Род-Айленд, США
[6] University of Utrecht, Утрехт, Нидерланды
[7] University of Sydney, Сидней, Новый Южный Уэльс, Австралия
[8] Beth Israel Deaconess Medical Center, Бостон, Массачусетс, США
[9] Radiology Editorial Office, Бостон, Массачусетс, США
[10] VU University Medical Center, Амстердам, Нидерланды

АННОТАЦИЯ

Диагностические исследования (*diagnostic accuracy studies*), как и другие клинические исследования, подвержены риску систематических ошибок (*bias*) из-за недостатков дизайна и проведения, а их результаты могут оказаться неприменимыми к другим группам пациентов и в других условиях. Читатели должны быть достаточно подробно проинформированы о дизайне и проведении диагностического исследования, чтобы судить о надёжности (*trustworthiness*) и применимости (*applicability*) его результатов. Руководство STARD (Standards for Reporting of Diagnostic Accuracy Studies) разработано с целью обеспечить полноту и прозрачность отчётов о диагностических исследованиях, содержит перечень основных пунктов отчёта, который может быть использован авторами, рецензентами и читателями как контрольный список (*checklist*) для отслеживания полноты представляемой информации.

Представлено обновлённое руководство **STARD, все материалы которого, включая контрольный список, доступны** на http://www.equator-network.org/reporting-guidelines/stard. Приведены обоснования для 30 пунктов руководства и описание того, что требуется от авторов для составления достаточно информативных отчётов об исследованиях.

Настоящая статья является русскоязычным переводом оригинальной публикации [Cohen JF, Korevaar DA, Altman DG, et al. STARD 2015 guidelines for reporting diagnostic accuracy studies: explanation and elaboration. *BMJ Open*. 2016;6:e012799. doi: 10.1136/bmjopen-2016-012799] под редакцией д.м.н. Р.Т. Сайгитова.

Ключевые слова: STARD; диагностические исследования; клинические исследования; дизайн исследования; систематические ошибки; надёжность; применимость.

Как цитировать

Cohen J.F., Korevaar D.A., Altman D.G., Bruns D.E., Gatsonis C.A., Hooft L., Irwig L., Levine D., Reitsma J.B., de Vet H.C.W., Bossuyt P.M.M. Рекомендации по составлению отчётов о диагностических исследованиях (STARD 2015): разъяснения и уточнения // *Digital Diagnostics*. 2021;2(3):313–342. DOI: https://doi.org/10.17816/DD71031

Рукопись получена: 15.01.2021 **Рукопись одобрена:** 18.05.2021 **Опубликована:** 26.05.2021

ЭКО•ВЕКТОР

DOI: https://doi.org/10.17816/DD71031

STARD 2015 guidelines for reporting diagnostic accuracy studies: explanation and elaboration. Translation to Russian

Jérémie F. Cohen[1, 2], Daniël A. Korevaar[1], Douglas G. Altman[3], David E. Bruns[4], Constantine A. Gatsonis[5], Lotty Hooft[6], Les Irwig[7], Deborah Levine[8, 9], Johannes B. Reitsma[6], Henrica C.W. de Vet[10], Patrick M. M. Bossuyt[1]

[1] University of Amsterdam, Amsterdam, The Netherlands
[2] Paris Descartes University, Paris, France
[3] University of Oxford, Oxford, UK
[4] University of Virginia School of Medicine, Charlottesville, Virginia, USA
[5] Brown University School of Public Health, Providence, Rhode Island, USA
[6] University of Utrecht, Utrecht, The Netherlands
[7]University of Sydney, Sydney, New South Wales, Australia
[8] Beth Israel Deaconess Medical Center, Boston, Massachusetts, USA
[9] Radiology Editorial Office, Boston, Massachusetts, USA
[10] VU University Medical Center, Amsterdam, The Netherlands

ABSTRACT

Diagnostic accuracy studies are, like other clinical studies, at risk of bias due to shortcomings in design and conduct, and the results of a diagnostic accuracy study may not apply to other patient groups and settings. Readers of study reports need to be informed about study design and conduct, in sufficient detail to judge the trustworthiness and applicability of the study findings. The STARD statement (Standards for Reporting of Diagnostic Accuracy Studies) was developed to improve the completeness and transparency of reports of diagnostic accuracy studies. STARD contains a list of essential items that can be used as a checklist, by authors, reviewers and other readers, to ensure that a report of a diagnostic accuracy study contains the necessary information. STARD was recently updated. All updated STARD materials, including the checklist, are available at http://www.equator-network.org/reporting-guidelines/stard. Here, we present the STARD 2015 explanation and elaboration document. Through commented examples of appropriate reporting, we clarify the rationale for each of the 30 items on the STARD 2015 checklist, and describe what is expected from authors in developing sufficiently informative study reports.

This article is the reprint with Russian translation edited by Dr. Ruslan Saygitov. The original that can be observed here: Cohen JF, Korevaar DA, Altman DG, et al. STARD 2015 guidelines for reporting diagnostic accuracy studies: explanation and elaboration. *BMJ Open* 2016;6:e012799. doi: 10.1136/bmjopen-2016-012799.

Keywords: STARD; diagnostic accuracy studies; clinical studies; bias; study design; applicability; trustworthiness.

To cite this article

Cohen JF, Korevaar DA, Altman DG, Bruns DE, Gatsonis CA, Hooft L, Irwig L, Levine D, Reitsma JB, de Vet HCW, Bossuyt PMM. STARD 2015 guidelines for reporting diagnostic accuracy studies: explanation and elaboration. Translation to Russian. *Digital Diagnostics*. 2021;2(3):313–342. DOI: https://doi.org/10.17816/DD71031

Received: 15.01.2021 **Accepted:** 18.05.2021 **Published:** 26.05.2021

ECO•VECTOR

DOI: https://doi.org/10.17816/DD71031

诊断测试报告指南 (STARD 2015)：澄清和澄清

Jérémie F. Cohen[1, 2], Daniël A. Korevaar[1], Douglas G. Altman[3], David E. Bruns[4], Constantine A. Gatsonis[5], Lotty Hooft[6], Les Irwig[7], Deborah Levine[8, 9], Johannes B. Reitsma[6], Henrica C.W. de Vet[10], Patrick M. M. Bossuyt[1]

[1] University of Amsterdam, Amsterdam, The Netherlands
[2] Paris Descartes University, Paris, France
[3] University of Oxford, Oxford, UK
[4] University of Virginia School of Medicine, Charlottesville, Virginia, USA
[5] Brown University School of Public Health, Providence, Rhode Island, USA
[6] University of Utrecht, Utrecht, The Netherlands
[7]University of Sydney, Sydney, New South Wales, Australia
[8] Beth Israel Deaconess Medical Center, Boston, Massachusetts, USA
[9] Radiology Editorial Office, Boston, Massachusetts, USA
[10] VU University Medical Center, Amsterdam, The Netherlands

简评

诊断研究(diagnostic accuracy studies)与其他临床试验一样，由于设计和行为的缺陷，存在系统错误 (bias) 的风险， 他们的结果可能不适用于其他患者群体和环境。 应充分详细地告知读者诊断研究的设计和实施， 判断可靠性(trustworthiness)，及其结果的适用性 (applicability)。 STARD 须知（Standards for Reporting of Diagnostic Accuracy Studies）旨在确保诊断研究报告的完整性和透明度，包含一份报告要点清单，可供作者、审稿人和读者用作检查清单 (checklist)以跟踪所提供信息的完整性。

已提供更新的 STARD 手册，所有手册（包括检查表）均可在以下网址获得 http://www.equator-network.org/reporting-guidelines/stard。提供 30 点指南的基本原理，并描述作者需要什么才能编制合理的信息丰富的研究报告。

这篇文章是原始出版物的俄语翻译 [Cohen JF, Korevaar DA, Altman DG, et al. STARD 2015 guidelines for reporting diagnostic accuracy studies: explanation and elaboration. BMJ Open. 2016;6:e012799. doi: 10.1136/bmjopen-2016-012799]由医学博士编辑R. T. Saygitova。

关键词：STARD； 诊断研究； 临床研究； 学习规划; 系统误差； 可靠性; 适用性。

引用本文

Cohen JF, Korevaar DA, Altman DG, Bruns DE, Gatsonis CA, Hooft L, Irwig L, Levine D, Reitsma JB, de Vet HCW, Bossuyt PMM. 诊断测试报告指南 (STARD 2015)：澄清和澄清. *Digital Diagnostics*. 2021;2(3):313–342. DOI: https://doi.org/10.17816/DD71031

收到： 15.01.2021 **接受：** 18.05.2021 **发布日期：** 26.05.2021

Список сокращений

КТ — компьютерная томография

КТК — КТ-колонография

МРТ — магнитно-резонансная томография

ЭКГ — электрокардиограмма

CONSORT (Consolidated Standards of Reporting Trials — *единые стандарты представления результатов испытаний*) — в заявлении представлены перечень вопросов и схема проведения рандомизированных контролируемых исследований, которые могут быть использованы авторами при составлении отчётов о результатах

QUADAS-2 (Quality Assessment Tool For Diagnostic Accuracy Studies) — опросник, разработанный для оценки качества диагностических исследований точности

STARD (Standards for Reporting of Diagnostic Accuracy Studies) — стандарты отчётности об исследованиях точности диагностики (http://www.equator-network.org/reporting-guidelines/stard)

ВВЕДЕНИЕ

Диагностические исследования (*diagnostic accuracy studies*) подвержены риску систематических ошибок (*bias*), что характерно и для других клинических исследований. Основные источники систематических ошибок кроются в методологических недостатках, особенностях отбора участников, сбора данных, выполнения или интерпретации результатов диагностического теста, анализа данных [1, 2]. В итоге показатели чувствительности (*sensitivity*) и специфичности (*specificity*) такого теста, сравниваемые с показателями референсного теста (*reference standard*), могут быть ошибочными, систематически отклоняясь от тех результатов, которые могли быть получены в идеальных условиях (табл. 1). Подобные смещения приводят к неверным рекомендациям по тестированию, негативно влияют на исходы пациентов и политику здравоохранения в целом.

Диагностическая точность[1] (*accuracy*) не является неотъемлемым свойством теста. При идентификации пациентов с изучаемым состоянием (*target condition*) точность теста зависит от условий его проведения, характеристик пациентов и результатов предыдущего тестирования [2]. Эти источники вариабельности диагностической точности актуальны для тех, кто пытается ответить на конкретный вопрос о применимости результатов исследования к определённым условиям. Риск систематических ошибок и опасения по поводу применимости результатов исследования — два ключевых компонента инструмента QUADAS-2, разработанного для оценки качества диагностических исследований [3].

Таблица 1. Основные термины руководства STARD

Термин	Значение
Медицинский тест	Любой метод сбора дополнительной информации о текущем или будущем состоянии здоровья пациента
Индексный (основной) тест (*index test*)	Исследуемый тест
Целевое состояние	Заболевание или состояние, которое, как ожидается, будет обнаружено с помощью индексного теста
Клинический референсный тест (*reference standard*)	Наилучший доступный метод для установления наличия или отсутствия целевого состояния. Безошибочный референсный стандарт — золотой стандарт
Чувствительность	Доля лиц с целевым состоянием и положительными результатами индексного теста
Специфичность	Доля лиц без целевого состояния и отрицательными результатами индексного теста
Предназначение теста	Использование индексного теста для диагностики, скрининга, определения стадии заболевания, мониторинга, надзора, предсказания, прогнозирования или других целей
Роль теста	Положение индексного теста по отношению к другим тестам при применении в одинаковых обстоятельствах: например, тест, используемый для сортировки (предварительный тест), замещающий тест, дополнительный тест или новый тест
Сомнительные результаты	Результаты, которые не являются положительными или отрицательными

[1] Здесь: соответствие референсному или истинному значению.

Читатели могут судить о риске систематической ошибки и применимости результатов диагностического исследования только в том случае, если они найдут необходимую для этого информацию в отчёте об исследовании. Опубликованный отчёт должен содержать всю важную информацию, на основании которой можно судить о надёжности (*trustworthiness*) и актуальности (*relevance*) выводов исследования вместе с полным и информативным описанием его результатов.

К сожалению, в нескольких обзорах было показано, что отчёты диагностических исследований зачастую непрозрачно описывают ключевые элементы [4–6]. Важная информация об участниках, дизайне исследования и фактических результатах часто отсутствует, а рекомендации авторов о применении изученного теста — часто чрезмерны и чересчур оптимистичны.

Руководство STARD имеет целью способствовать составлению более полных и прозрачных отчётов о диагностических исследованиях [7]. По аналогии со стандартами представления результатов рандомизированных исследований (CONSORT) [8, 9] рекомендации STARD представлены в виде перечня пунктов, которые следует описывать в любых отчётах о диагностических исследованиях.

Руководство STARD впервые опубликовано в 2003 г. и пересмотрено в 2015 г. [10]. Обновление было выполнено для того, чтобы дополнить руководство актуальными сведениями об источниках систематической ошибки, вариабельности и других проблемах отчётности, а также упростить использование рекомендаций STARD. Обновлённое руководство теперь включает 30 ключевых пунктов (табл. 2).

Таблица 2. Контрольный перечень STARD 2015 [10]

Раздел статьи	№	Пункт
Название или аннотация	1	Обозначение (в названии или в аннотации) как исследования диагностической точности (диагностическое исследование) с указанием хотя бы одного показателя точности (таких как чувствительность, специфичность, прогностическая значимость или площадь под ROC-кривой)
Аннотация	2	Структурированное краткое описание дизайна исследования, методов, результатов и выводов (более подробно см. рекомендации STARD по оформлению аннотаций)
Введение	3	Научные и клинические предпосылки, включая предполагаемое использование и клиническую роль индексного теста
	4	Цели и гипотезы исследования
Методы		
Дизайн исследования	5	Источники и сбор данных до (проспективное исследование) или после (ретроспективное исследование) проведения индексного и референсного тестов
Участники	6	Критерии отбора
	7	Основания, по которым определяли участников, потенциально подходящих для отбора в исследование (симптомы, результаты предыдущих диагностических тестов, включение в регистр)
	8	Место и время скрининга участников, потенциально подходящих для отбора в исследование (учреждения, местоположение и даты)
	9	Формирование выборки участников: последовательная, случайная или «удобная» для исследователя
Методы диагностического исследования	10а	Детальное описание индексного теста, позволяющее повторить его
	10б	Детальное описание референсного теста, позволяющее повторить его
	11	Обоснование выбора референсного теста (при наличии аналогов)
	12а	Определение и обоснование пороговых значений положительных результатов или категорий индексного теста, различая запланированное и выведенное в результате разведочного анализа
	12б	Определение и обоснование пороговых значений положительных результатов или категорий референсного теста, различая запланированное и выведенное в результате разведочного анализа
	13а	Доступность клинических данных и информации о результатах референсного теста тем, кто проводил или фиксировал результаты индексного теста
	13б	Доступность клинических данных и информации о результатах индексного теста тем, кто оценивал результаты референсного теста

Таблица 2. Окончание

Раздел статьи	№	Пункт
Анализ	14	Методы оценки или сравнения показателей диагностической точности
	15	Действия в отношении сомнительных результатов индексного или референсного тестов
	16	Действия в отношении отсутствующих (неполных) данных индексного и референсного тестов
	17	Анализ вариабельности диагностической точности с дифференциацией запланированного и установленного после получения данных результатов разведочного анализа
	18	Запланированный размер выборки и его определение
Результаты		
Участники	19	Формирование выборки исследования
	20	Характеристика участников исследования (демографические и клинические данные)
	21a	Распределение пациентов с целевым состоянием по тяжести заболевания
	216	Распределение пациентов без целевого состояния по альтернативным диагнозам
Результаты диагностического исследования	22	Временной интервал и любые медицинские вмешательства между выполнением индексного и референсного тестов
	23	Таблицы сопряжённости (или распределение) результатов индексного и референсного тестов
	24	Оценка диагностического показателя и её точность (например, 95% доверительный интервал)
	25	Любые нежелательные последствия выполнения индексного или референсного тестов
Обсуждение	26	Ограничения исследования, включая источники потенциальных систематических ошибок, статистической неопределённости и ограниченной обобщаемости результатов
	27	Значение для практики, включая предполагаемое использование и клиническую роль индексного теста
Дополнительная информация	28	Регистрационный номер исследования и наименование регистра
	29	Доступ к полному протоколу исследования
	30	Источники финансирования, другие виды поддержки и роль спонсоров исследования

Ниже представлены рекомендации STARD 2015 с пояснениями и уточнениями. Это расширенная и обновлённая версия документа, опубликованного в 2003 г. [11]. Комментируя в качестве примеров фрагменты из опубликованных работ, мы обосновываем применение каждого пункта руководства STARD 2015 и описываем, что ожидается от авторов.

Мы уверены, что представленная нами информация поможет исследователям в написании информативных исследовательских отчётов, а также поможет рецензентам, редакторам и читателям убедиться, что представленные на рассмотрение и опубликованные рукописи о диагностических исследованиях достаточно подробны.

КОНТРОЛЬНЫЙ ПЕРЕЧЕНЬ STARD 2015: ПОЯСНЕНИЯ И УТОЧНЕНИЯ

Название или аннотация

Пункт 1. ***Обозначение (в названии или в аннотации) как исследования диагностической точности (диагностическое исследование) с указанием хотя бы одного показателя точности, таких как чувствительность, специфичность, прогностическая значимость (predictive values) или площадь под ROC-кривой***

Пример

«Основные показатели исходов: чувствительность и специфичность КТ-колонографии при выявлении лиц с прогрессирующей неоплазией (прогрессирующая аденома или колоректальный рак) с диаметром новообразования не менее 6 мм» [12].

Пояснение

Электронные базы данных, такие как MEDLINE и Embase, незаменимы при поиске биомедицинских исследований по определённой теме. Чтобы облегчить поиск своей статьи, авторы могут явно идентифицировать её как отчёт о диагностическом исследовании. Это может быть выполнено с использованием в заголовке и/или аннотации терминов, которые относятся к показателям диагностической точности, например «чувствительность» (*sensitivity*), «специфичность» (*specificity*), «положительная прогностическая значимость» (*positive predictive value*), «отрицательная прогностическая

значимость» (*negative predictive value*), «площадь под ROC-кривой» (*area under the curve, AUC*) или «отношение правдоподобия» (*likelihood ratio*).

С 1991 г. в MEDLINE для индексирования диагностических исследований введено специальное ключевое слово (заголовок предметной рубрики MeSH, *Medical Subject Headings*) «Чувствительность и Специфичность» (Sensitivity and Specificity). К сожалению, чувствительность поиска таких исследований по заголовку MeSH не превышает 51% [13]. По состоянию на май 2015 г., в словаре Embase (тезаурус Emtree) содержатся ключевые слова (теги) для 38 типов исследований: «исследование точности диагностических тестов» (*diagnostic test accuracy study*) — одно из них, но появилось оно лишь в 2011 г.

В приведённом выше примере авторы упомянули термины «чувствительность» и «специфичность» в аннотации. При использовании одного из этих терминов в поисковом запросе статья будет извлечена из базы данных и легко идентифицирована как диагностическое исследование.

Аннотация

Пункт 2. ***Структурированное краткое описание дизайна исследования, методов, результатов и выводов (более подробно см. рекомендации STARD по оформлению аннотаций)***

Пример

См. рекомендации STARD по оформлению аннотаций (*контрольный перечень доступен на* https://www.equator-network.org/reporting-guidelines/stard-abstracts/)

Пояснение

Читатели используют аннотации, чтобы решить, следует ли им открыть полный отчёт об исследовании и потратить время на его чтение. В тех случаях, когда невозможно получить доступ к полному отчёту об исследовании или когда время ограничено, можно предположить, что клинические решения будут основываться только на информации, представленной в аннотации.

В двух недавних литературных обзорах аннотации диагностических исследований, опубликованные в журналах с высоким импакт-фактором или представленные на международной научной конференции, были признаны недостаточно информативными, поскольку ключевые данные о цели исследования, его методах, результатах и их применимости часто отсутствовали [14, 15].

Информативные аннотации помогают читателям оперативно и критично оценивать достоверность (*validity*) исследования (риск систематических ошибок, *risk of bias*) и применимость его результатов к клиническим условиям (обобщаемость, *generalisability*). Структурированные аннотации с отдельными заголовками для целей, методов, результатов и их интерпретации упрощают читателям поиск необходимой информации [16].

Основанные на STARD 2015 недавно разработанные рекомендации STARD для аннотаций (STARD for Abstracts) содержат ключевые пункты, которые должны быть включены в аннотации журнальных статей или материалов конференций.

Введение

Пункт 3. ***Научные и клинические предпосылки, включая предполагаемое использование и клиническую роль индексного теста***

Пример

«Необходимость повышения эффективности использования рентгенографии в отделениях неотложной помощи уже давно подтверждена документально. Такая потребность часто возникает в отношении пациентов с острой травмой голеностопного сустава, которых обычно направляют на рентгенографию, несмотря на то, что вероятность перелома составляет менее 15%. Процедура направления и результаты рентгенографии для пациентов с травмами коленного сустава описаны менее чётко и могут быть менее эффективными, чем для пациентов с травмами голеностопа <...>. Огромный объём недорогих тестов, таких как обычная рентгенография, может способствовать росту затрат на здравоохранение в такой же степени, как и высокотехнологичные процедуры, проводимые в небольшом количестве. <...>. Если это будет подтверждено в последующих исследованиях, правило принятия решения для пациентов с травмой колена может привести к значительному сокращению использования рентгенографии колена и значительной экономии средств здравоохранения без ущерба для пациента» [17].

Пояснение

Во введении к отчётам о научных исследованиях авторы должны обосновать необходимость их проведения. При этом они могут ссылаться на предыдущие работы по теме, сохраняющуюся неопределённость и клинические последствия этих пробелов в знаниях (*knowledge gap*). Чтобы помочь читателям оценить значение исследования, авторы могут разъяснить предполагаемое использование и клиническую роль изучаемого теста (*index test*).

Тест может быть предназначен для таких целей, как диагностика, скрининг, определение стадии заболевания (*staging*), мониторинг, надзор (*surveillance*), прогнозирование, выбор терапии или другое [18]. Клиническая роль изучаемого теста связана с его ожидаемой позицией относительно других тестов в клиническом протоколе (*clinical pathway*) [19]. Например, предварительный тест (*triage test*) будет использоваться перед существующим тестом, потому что он менее затратный или обременительный, но часто также и менее точен. Дополнительный тест (*add-on test*) будет проводиться после существующих тестов для повышения точности общей стратегии тестирования путём выявления

ложноположительных или ложноотрицательных результатов первоначального теста. В некоторых случаях вместо основного теста может использоваться новый.

Определение предназначения и клинической роли теста определит дизайн (схему) исследования, а также целевой уровень чувствительности и специфичности; из этих определений следуют критерии отбора, как и где искать подходящих участников, как выполнять тесты и интерпретировать их результаты [19].

Определение клинической роли полезно для оценки относительной значимости потенциальных ошибок (ложноположительных и ложноотрицательных результатов), допущенных при выполнении исследуемого или индексного теста (*index test*). Например, предварительный тест для исключения заболевания должен быть высокочувствительным, тогда как тест, нацеленный на выявление заболевания, — высокоспецифичным.

В вышеприведённом примере предполагаемая цель использования — диагностика переломов у пациентов с острыми травмами коленного сустава, а потенциальная клиническая роль — предварительный тест с целью сортировки пациентов. Рентгенография (основной тест) будет проводиться лишь у пациентов с положительным результатом недавно разработанного правила принятия решения. Авторам следует описать современные научные и клинические предпосылки изучаемой проблемы со здоровьем, а также причину, в связи с которой они стремятся разработать предварительный тест: сокращение количества рентгенографических исследований и, как следствие, расходов на медицинское обслуживание.

Пункт 4. ***Цели и гипотезы исследования***

Пример 1

Цель исследования — оценить чувствительность и специфичность трёх различных диагностических стратегий: однократный экспресс-тест на антиген, экспресс-тест на антиген с повторным экспресс-тестом в случае отрицательного результата (стратегия тест–тест) и экспресс-тест на антиген с последующим посевом в случае отрицательного результата (стратегия тест–посев, предложенная Американской академией педиатрии). Все полученные результаты сравнивали с золотым стандартом — культивированием в двух чашках. Кроме того, <...> сравнили способность этих стратегий достигать абсолютную чувствительность диагностического теста >95%» [20].

Пример 2

«Наши основные гипотезы: 1) экспресс-тесты на антиген, выполняемые в кабинете врача, более чувствительны, чем посевы в чашках с кровяным агаром, выполненные и оцениваемые там же, когда каждый тест сравнивается с результатами одновременно проведённого и интерпретируемого посева в чашки с кровяным агаром в больничной лаборатории; 2) чувствительность экспресс-теста на антиген подвержена систематической ошибке, связанной с неоднородностью целевой популяции» [21].

Пояснение

Клинические исследования могут иметь общую цель (долгосрочную, например «добиться снижения стадии рака пищевода»), конкретные задачи (чётко определённые цели для данного исследования) и проверяемые гипотезы (утверждения, которые могут быть опровергнуты результатами исследования).

В диагностических исследованиях статистические гипотезы, как правило, выдвигаются в терминах критериев приемлемости (качества) для отдельных тестов (минимальные уровни чувствительности, специфичности или других показателей). В этих случаях гипотезы обычно содержат количественное выражение ожидаемого значения диагностического параметра. В других случаях статистические гипотезы могут быть сформулированы в терминах эквивалентности (*equality*) или не меньшей точности (*non inferiority in accuracy*) при сравнении двух или более индексных тестов.

Предварительное описание гипотез исследования ограничивает риски, связанные с апостериорным (незапланированным) анализом данных (*data-dredging*) и ложными находками, поспешными выводами о выполнимости тестов или субъективными суждениями об их точности. Цели и гипотезы также необходимы при расчётах размера выборки. Обзор 126 отчётов о диагностических исследованиях, опубликованных в журналах с высоким импакт-фактором в 2010 г., показал, что 88% из них не содержали чётко сформулированных гипотез [22].

Выше, в первом примере, целью авторов была оценка точности трёх диагностических стратегий. Конкретная гипотеза заключалась в том, что чувствительность любой из стратегий превысит заранее установленное значение 95%. Во втором примере авторы чётко описывают гипотезы, которые они планируют проверить в своём исследовании. Первая гипотеза — о сравнении чувствительности двух индексных тестов, выполняемых в кабинете врача (экспресс-тест на антигены и посев); вторая — о вариабельности результатов экспресс-теста в зависимости от характеристик пациента (*spectrum bias*).

Методы

Пункт 5. ***Источники и сбор данных до (проспективное исследование) или после (ретроспективное исследование) проведения индексного и референсного тестов***

Пример

«Изучили базу данных пациентов, прошедших процедуру тонкоигольной локализации новообразований и их удаления хирургическим путём с помощью цифрового томосинтеза молочной железы в период с апреля 2011 г. по январь 2013 г. <...> Затем медицинские карты пациентов и изображения 36 выявленных поражений

были ретроспективно просмотрены автором более чем с 5-летним опытом лучевых исследований молочных желёз после прохождения соответствующей программы стажировки» [23].

Пояснение

На сегодняшний день термины «проспективный» и «ретроспективный» не имеют чёткого определения, поэтому авторам необходимо чётко описать, планировался ли сбор данных до или после проведения индексного *(index test)* и референсного (*reference standard*) тестов. Если авторы определили вопрос исследования до проведения индексного и референсного тестов, они могут предпринять соответствующие действия для оптимизации процедур в соответствии с протоколом исследования и для сбора необходимых данных [24].

Иногда идея исследования возникает после получения результатов тестирования, представляющего исследовательский интерес. В таких случаях необходимые данные извлекают из медицинских карт пациентов или регистров. Ретроспективные исследования могут лучше отражать обычную клиническую практику, чем проспективные, но при этом исследователи могут идентифицировать не всех пациентов, соответствующих критериям отбора, и получить данные низкого качества с большим количеством пропусков (*missing data*) [24]. Причиной этого может быть, например, то, что в повседневной медицинской практике не все пациенты, прошедшие интересующее исследователей тестирование, будут протестированы в том числе и с применением референсного теста.

В примере выше данные были явно собраны ретроспективно: участников идентифицировали путём скрининга базы данных, клинические сведения извлекали из медицинских карт пациентов, хотя снимки интерпретировали заново.

Пункт 6. ***Критерии отбора***

Пример 1

«Подходящими для включения в исследование были взрослые (старше 18 лет) с подозрением на тромбоэмболию лёгочной артерии на основании наличия хотя бы одного из следующих симптомов: необъяснимая (внезапная) одышка, ухудшение имеющейся одышки, боль при вдохе или необъяснимый кашель. Мы не включали пациентов, получавших антикоагулянтную терапию (антагонисты витамина К или гепарин) на момент первичного обследования, беременных, при невозможности последующего наблюдения, а также пациентов, которые не хотели или не могли предоставить письменное информированное согласие» [25].

Пример 2

«Для участия в исследовании отбирали пациентов («случаи») с признаками диареи, при обнаружении токсина методом иммуноферментного анализа и токсигенного штамма *C. difficile* при посеве (в образце, взятом менее чем за 7 суток до выявления штамма). Определяли диарею как неоформленный или жидкий стул три и более раз в день. В исследование не включали детей и взрослых в отделениях интенсивной терапии или гематологии. Подходили также пациенты с первым рецидивом после завершения лечения предыдущей инфекции *C. difficile*, но не пациенты с последующими рецидивами. <...> К каждому «случаю» подбирали по 9 пациентов контрольной группы. Эти пациенты находились в той же палате, и находились в непосредственной близости пациентов группы «случай». «Контроли» не имели признаков диареи либо имели таковые, но в сочетании с отрицательным результатом иммуноферментного анализа и посева (в образце, взятом менее 7 суток до тестирования)» [26].

Пояснение

Поскольку диагностическое исследование описывает действие теста при определённых обстоятельствах, отчёт об исследовании должен включать полное описание критериев, которые использовались для определения подходящих участников. Критерии отбора (*eligibility criteria*), как правило, связаны с характером и стадией исследуемого или целевого состояния (*target condition*) и предполагаемым применением результатов индексного теста в будущем. Они часто включают признаки, симптомы или результаты предыдущих тестов, которые вызывают определённые подозрения относительно наличия целевого состояния. Для невключения или исключения участников по соображениям безопасности, практической осуществимости и этики могут использоваться дополнительные критерии.

Невключение пациентов с определённым заболеванием или получающих определённое лечение, которое, как известно, отрицательно влияет на результаты теста, может привести к завышенным оценкам диагностической точности [27]. В качестве примера можно привести пациентов, получающих β-блокаторы в исследованиях, где оценивается диагностическая точность электрокардиограммы (ЭКГ) с физической нагрузкой.

Некоторые исследования имеют одну группу критериев отбора для всех участников; их иногда называют одновыборочные (*single-gate)* или когортные исследования (*cohort studies*). В других исследованиях одна группа критериев отбора применяется к участникам с целевым состоянием, другая — к участникам без такового; такие исследования называют многовыборочными (*multiple-gate)* или исследованиями «случай-контроль» (*case-control studies*) [28].

В первом примере, представленном выше, критерии отбора представляют признаки и симптомы, возрастные ограничения и критерии невключения, соответствующие определённым состояниям и методам лечения. Поскольку ко всем участникам исследования применяются одинаковые критерии отбора, речь идёт об одновыборочном исследовании.

Во втором примере авторы применяли разные критерии отбора к участникам с целевым состоянием и без него: одна группа состояла из пациентов с подтверждённым диагнозом *Clostridium difficile*-ассоциированной инфекции, другая включала здоровых «контролей». Это пример многовыборочного исследования. Значительные различия между тяжёлыми «случаями» и здоровыми «контролями» могут привести к завышенным оценкам точности теста [6, 29].

Пункт 7. ***Основания, по которым определяли участников, потенциально подходящих для отбора в исследование (симптомы, результаты предыдущих диагностических тестов, включение в регистр)***

Пример

«Изучили базу данных пациентов, прошедших процедуру тонкоигольной локализации новообразований и их удаления хирургическим путём с помощью цифрового томосинтеза молочной железы в период с апреля 2011 г. по январь 2013 г.» [23].

Пояснение

Критерии отбора определяют лиц, которые могут участвовать в исследовании, однако они не описывают, как авторы исследования определили подходящих участников. Подбор участников осуществляют разными способами [30]. Врач общей практики может в рабочее время оценивать каждого пациента на соответствие критериям отбора. Исследователи могут извлекать данные потенциальных участников из регистров отделений неотложной помощи. В одних исследованиях пациентов идентифицируют только после прохождения индексного тестирования, в других — после выполнения референсного теста. Многие ретроспективные исследования включают участников, упоминаемых в больничных базах данных, при условии выполнения обоих тестов — индексного и референсного [31].

Различия в методах выявления пациентов, соответствующих критериям отбора, могут влиять на характеристики и распространённость целевого состояния в исследуемой группе, а также на диапазон и относительную частоту альтернативных (сопутствующих) состояний у пациентов без него [32]. Всё это может отражаться на оценках диагностической точности.

В примере выше участников отбирали из базы данных пациентов при условии наличия данных о проведении индексного (цифровой томосинтез) и референсного (маммографическое исследование) тестов.

Пункт 8. ***Место и время скрининга участников, потенциально подходящих для отбора в исследование (учреждения, местоположение и даты)***

Пример

«Исследование было проведено в отделении неотложной помощи детской больницы при университете в период с 21 января 1996 г. по 30 апреля 1996 г.» [33].

Пояснение

Результаты диагностического исследования отражают конкретный клинический контекст и условия выполнения теста (*setting*). Например, медицинский тест может выполняться по-разному в условиях учреждений первичной, вторичной или третичной медицинской помощи, поэтому авторы должны описать фактические условия, в которых проводилось исследование, а также указать точное местоположение: названия участвующих медицинских центров, города и страны. Спектр (разнообразие характеристик) целевого состояния, а также диапазон других состояний, которые возникают у пациентов с подозрением на целевое состояние, могут варьировать в зависимости от условий проведения исследования и механизмов направления пациентов за помощью (*referral **mechanisms***) [34–36].

Поскольку процедуры тестирования, механизмы направления к специалистам, а также распространённость и степень тяжести заболеваний могут меняться со временем, авторы должны сообщать даты начала и окончания набора участников.

Эта информация существенна для читателей, желающих оценить обобщаемость (*generalisability*) результатов исследования и их применимость к определённым вопросам, а также для тех, кто хотел бы использовать полученные в ходе исследования свидетельства (*evidence*) для принятия обоснованных решений в области здравоохранения.

В приведённом выше примере чётко описаны условия и указаны даты проведения исследования.

Пункт 9. ***Формирование выборки участников: последовательная, случайная или «удобная» для исследователя***

Пример

«Первый автор (E.N.E.) оценивал и подбирал участников исследования в соответствии с критериями отбора до включения в исследование. Это была «удобная» для наблюдения исследователем выборка (*convenience sample*) детей с фарингитом. Набор испытуемых был произведён в период пребывания первого автора в отделении неотложной помощи» [37].

Пояснение

Включённые в исследование участники могут составлять либо последовательную выборку всех пациентов (*consecutive series*), оцениваемых на соответствие критериям отбора и удовлетворяющих критериям включения, либо их ограниченное количество. Ограниченная выборка может быть полностью случайной, сформированной на основе таблицы случайных чисел, или не случайной, если пациентов набирают только в определённые дни или в определённые часы работы. В последнем случае включённые участники не могут считаться репрезентативной выборкой целевой популяции (*targeted*

population), а обобщение результатов исследования может иметь ограничения [2, 29].

В примере выше авторы подробно описали «удобную» выборку (*convenience series*), в которую участников отбирали, основываясь на их доступности для исследователя.

Пункт 10a. ***Детальное описание индексного теста, позволяющее повторить его***

Пункт 10б. ***Детальное описание референсного теста, позволяющее повторить его***

Пример

«Внутривенный катетер вводили в срединную локтевую вену, образцы крови собирали в пробирки до стресс-теста (исходные данные), сразу после него и через 1,5 и 4,5 часа по его завершении. Взятые образцы крови после сбора помещали на лёд на 1 час и в последующем до проведения анализа хранили при температуре -80°C. Перед анализом на определение сердечного тропонина I (cTnI) допускалось однократное размораживание/замораживание проб. Концентрацию высокочувствительного **cTnI измеряли прототипом ме**тода высокочувствительного анализа (ARCHITECT *STAT* high-**sensitivity troponin, Abbott Diagnostics), где с помо**щью иммобилизованных антител распознавали эпитопы 24–40 и посредством детектирующих антител — эпитопы 41–49 cTnI. Предел обнаружения (*limit **of detec**tion*) для высокочувствительного анализа cTnI недавно установлен другими группами исследователей и составил 1,2 нг/л (16 нг/л — 99-й процентиль), а с учётом коэффициента вариации 10% — 3,0 нг/л. <...> Образцам с концентрациями **cTnI ниже указанного предела при**сваивали значение 1,2» [38].

Пояснение

Различия в выполнении индексного и референсного тестов — потенциальный источник вариаций диагностической точности [39, 40]. Именно поэтому авторы должны описывать методы выполнения индексного и референсного тестов достаточно подробно, чтобы позволить другим исследователям повторить исследование, а читателям — оценить (1) выполнимость индексного теста в своих условиях работы (*setting*), (2) адекватность референсного теста и (3) применимость результатов исследования к их клиническому вопросу. При этом описание должно охватывать ключевые элементы протокола тестирования, включая следующие:

1) преаналитическая фаза: например, подготовка пациентов (голодание/питание) перед забором крови, обработка образца до начала тестирования и связанные с этим ограничения (такие как нестабильность проб) или анатомическое расположение выполняемого измерения;
2) аналитическая фаза, включая используемые материалы, инструменты, аналитические процедуры (последовательность действий);
3) постаналитическая фаза: например, оценки риска по результатам анализа или другим переменным.

Различия между исследованиями в показателях точности теста, обусловленные различиями в протоколах тестирования, неоднократно описаны, включая, например, использование гипервентиляции перед проведением ЭКГ с физической нагрузкой и использование томографии для нагрузочной сцинтиграфии миокарда с таллием [27, 40].

Количество, профессиональная подготовка и компетентность лиц, выполняющих и интерпретирующих результаты индексного и референсного тестов, также могут иметь решающее значение. Во многих исследованиях показана вариабельность результатов тестирования, особенно при применении методов визуализации, в зависимости от квалификации интерпретирующих их лиц [41, 42]. Показано также, что качество анализа результатов цитологических и микробиологических исследований зависит от профессионального опыта, компетентности и предварительного обучения с целью повышения точности оценок и снижения расхождений в оценках между наблюдателями [43–45]. Информация об уровне подготовки специалистов, осуществляющих оценку и интерпретацию результатов тестирования, может помочь читателям сделать вывод о достижимости аналогичных результатов в условиях их деятельности.

В некоторых случаях исследование включает проведение нескольких референсных тестов. Например, пациенты с нарушениями, обнаруженными изучаемым методом визуализации (индексный тест), могут проходить процедуру биопсии с установлением окончательного диагноза по результатам гистологического исследования, тогда как клиническое наблюдение пациентов, у которых такие нарушения не обнаружены, будет включать референсный тест. Это может быть потенциальным источником систематической ошибки, поэтому авторам следует указать, какие группы пациентов какой референсный тест получили [2, 3].

В будущих специальных версиях STARD будет разработано более конкретное руководство по специализированным областям тестирования или определённым типам тестов. Эти рекомендации будут доступны на странице STARD вебсайта EQUATOR (Повышение качества и прозрачности исследований по вопросам здоровья; http://www.equator-network.org/).

В примере выше авторы описали, как отбирали и обрабатывали образцы крови в лаборатории. Они также сообщили аналитические характеристики индексного теста, полученные в предыдущих исследованиях.

Пункт 11. ***Обоснование выбора референсного теста (при наличии аналогов)***

Пример

«Международный нейропсихиатрический опросник MINI разработан для быстрого и эффективного

диагностического интервью как в научных целях, так и в клинической практике (*авторами приведена ссылка в поддержку данного утверждения*). Он считается более надёжным (*reliability rates*) и достоверным (*validity rates*) по сравнению с другими стандартными тестами, такими как SCID (Структурированное клиническое интервью для выявления психических нарушений) и CIDI (Структурированный международный диагностический опросник) (*авторами приведены ссылки в поддержку данного утверждения*)» [46].

Пояснение

В диагностических исследованиях референсный тест используется для установления наличия или отсутствия целевого состояния у участников исследования. Для определения одного и того же целевого состояния могут быть доступны несколько референсных тестов. В таком случае авторы должны обосновать свой выбор конкретного референсного теста из имеющихся альтернативных вариантов. Выбор может зависеть от предназначения (цели использования) индексного теста, клинической значимости, практических и/или этических соображений.

Альтернативные референсные тесты не всегда полностью согласуются друг с другом. Некоторые референсные тесты менее точны, чем другие. В других случаях референсные тесты отражают связанные, но разные проявления или стадии болезни, как в случае подтверждения болезни методом визуализации (первый референсный метод) в сравнении с диагностикой на основании клинически значимых событий (второй референсный метод).

В примере выше авторы выбрали MINI, структурированное диагностическое интервью, широко используемое в целях психиатрического освидетельствования, в качестве референсного инструмента для выявления признаков депрессии и риска самоубийства у взрослых с эпилепсией. Свой выбор авторы обосновали краткостью опросника, эффективностью при использовании как с клинической, так и научной целями, надёжностью и достоверностью по сравнению с альтернативными диагностическими опросниками.

Пункт 12a. ***Определение и обоснование пороговых значений положительных результатов или категорий индексного теста, различая запланированное и выведенное в результате разведочного анализа***

Пункт 12б. ***Определение и обоснование пороговых значений положительных результатов или категорий референсного теста, различая запланированное и выведенное в результате разведочного анализа***

Пример

«Мы также сравнили чувствительность модели риска при специфичности, соответствующей фиксированному пороговому значению положительного иммунохимического теста кала, составляющего 50 нг/мл. Мы использовали это пороговое значение, так как на момент исследования предполагалось его использование в голландской программе скрининга» [47].

Пояснение

Результаты тестов в их первоначальном виде могут быть дихотомическими (положительные или отрицательные), иметь несколько категорий (например, высокий, средний или низкий уровень риска) или быть непрерывными (интервал или шкала отношений).

Для тестов с несколькими категориями или непрерывными результатами конечный результат тестирования часто реклассифицируют в положительный (подтверждение заболевания) или отрицательный (исключение заболевания). Для этого необходимо определить критерии положительного результата теста: результаты, превышающие пороговое значение, будут считаться положительными результатами индексного теста. В некоторых исследованиях строят график ROC-кривой путём расчёта пар чувствительность–специфичность для всех возможных пороговых значений.

Чтобы оценить достоверность (*validity*) и применимость (*applicability*) этих классификаций, читателям необходимо знать критерии положительного результата или категорий результатов, как они были определены, и были ли они определены до начала исследования или после сбора данных. Запланированные пороговые значения (*prespecified thresholds*) могут быть основаны на (1) результатах предыдущих исследований, (2) пороговых значениях, используемых в клинической практике, (3) указанных в клинических рекомендациях или (4) рекомендованных производителем. Если пороговые значения ранее не были установлены, у авторов может возникнуть соблазн определить точность различных пороговых значений после сбора данных.

Если авторы определяют критерий положительного результата после проведения теста, выбирая тот, который максимизирует характеристики теста, существует высокий риск того, что полученные оценки точности теста будут слишком оптимистичными, особенно в небольших исследованиях [48, 49]. Последующие исследования могут не воспроизвести полученные результаты [50, 51].

В примере выше авторы обосновали свой выбор пороговых значений.

Пункт 13a. ***Доступность клинических данных и информации о результатах референсного теста тем, кто проводил или фиксировал результаты индексного теста***

Пункт 13б. ***Доступность клинических данных и информации о результатах индексного теста тем, кто оценивал результаты референсного теста***

Пример

«Снимки каждого пациента описывали два врача-рентгенолога, прошедшие обучение в области лучевых исследований мочеполовой системы и с опытом работы 12 и 8 лет соответственно, не имевшие доступа (*blinded*)

к данным пациента, включая окончательный гистологический диагноз» [52].

Пояснение

Некоторые медицинские тесты, в частности большинство методов визуализации, требуют участия человека в проведении, интерпретации и принятии решений. На эти действия может повлиять информация, которая доступна лицу, проводящему тестирование [1, 53, 54], что в свою очередь может привести к искусственно завышенной согласованности между тестами или между результатами индексного и референсного тестов.

Если у специалиста, выполняющего диагностику, есть доступ к информации о признаках, симптомах заболевания и результатах предыдущих тестов, это может привести к предвзятости в интерпретации, но в то же время может отражать применение теста в обычной клинической практике [2]. Напротив, при отсутствии достаточной информации для правильной интерпретации результатов индексного теста эффективность теста (*test performance*) может снизиться, а результаты исследования могут иметь ограниченную применимость. В любом случае читатели отчёта исследования должны знать, какая дополнительная информация была доступна исследователям-аналитикам и могла повлиять на их окончательные решения.

В других ситуациях специалистам, оценивающим референсный тест, могут быть известны результаты индексного теста. В таких случаях окончательная классификация может основываться на результатах индексного теста, а значит, представленные оценки точности индексного теста будут завышенными [1, 2, 27]. Тесты, требующие субъективной интерпретации, особенно подвержены таким систематическим ошибкам (*bias*).

Ограничение доступа исследователей-аналитиков к информации обычно называют «ослеплением» или «маскированием». Суть этого пункта руководства заключается не в том, чтобы обосновать преимущества или недостатки метода «ослепления», а в том, чтобы пояснить, что читателям отчёта об исследовании необходима информация о сокрытии (или несокрытии) сведений об индексном и референсном тестах, что позволит верно интерпретировать результаты исследования.

В примере выше специалисты, интерпретирующие результаты бесконтрастной компьютерной томографии с целью дифференциации ангиомиолипомы почек и почечно-клеточной карциномы, не имели доступа ни к клиническим данным, ни к результатам гистологического исследования, являвшегося в этом исследовании референсным методом диагностики.

Пункт 14. ***Методы оценки или сравнения показателей диагностической точности***

Пример

«Статистическое сравнение чувствительности и специфичности выполнено с помощью теста McNemar для зависимых (коррелируемых) признаков. Все тесты были двусторонними, проверялась гипотеза о том, что диагностические характеристики стереоскопической цифровой маммографии и цифровой маммографии отличаются. Статистически значимыми считали результаты при p <0,05» [55].

Пояснение

Для описания эффективности медицинского теста используется множество показателей диагностической точности, вычисление которых на основе собранных данных может вызывать трудности [56]. Авторы должны сообщить о методах вычисления показателей, которые они сочли подходящими для целей своего исследования.

Статистические методы могут быть использованы для проверки конкретных гипотез, вытекающих из целей исследования. В исследованиях с одним тестом авторы могут захотеть оценить, превышает ли диагностическая точность тестов предварительно установленный уровень (например, чувствительность не менее 95%, см. Пункт 4).

В диагностических исследованиях могут также сравнивать два и более индексных теста. В таких случаях проверка статистических гипотез обычно предполагает оценку превосходства одного теста над другим либо его не меньшей эффективности (*non-inferiority*) [57]. Для таких сравнений авторы должны указать меру различия, исходя из целей исследования, цели и роли индексного теста применительно к действующим клиническим рекомендациям. Примерами являются относительная чувствительность (*relative sensitivity*), абсолютный прирост чувствительности и относительное диагностическое отношение шансов (*relative diagnostic odds ratio*) [58].

В приведённом выше примере авторы использовали статистику теста McNemar для оценки различий чувствительности и специфичности стереоскопической и стандартной цифровой маммографии у пациентов с повышенным риском развития рака молочной железы. Величина p сама по себе не является количественным выражением относительной точности двух исследованных тестов. На значение p, как правило, влияют величина эффекта (различие между тестами) и размер выборки. В этом примере авторы могли рассчитать относительную или абсолютную разницу в чувствительности и специфичности, включая 95% доверительный интервал с учётом парного (связанного) характера данных.

Пункт 15. ***Действия в отношении сомнительных результатов индексного или референсного тестов***

Пример

«Сомнительные результаты считали ложноположительными или ложноотрицательными и включали в окончательный анализ. Например, сомнительный результат у пациента с аппендицитом считался отрицательным» [59].

Пояснение

Сомнительные результаты — это те, которые не являются ни положительными, ни отрицательными. Такие результаты могут быть получены при выполнении как индексного, так и референсного тестов, и являются проблемой при оценке их эффективности [60–63]. Частота сомнительных результатов варьирует от теста к тесту, в некоторых случаях доля таких результатов может достигать 40% [62].

Причин этому множество [62, 63]. Тест может быть неудачным по техническим причинам или вследствие недостатков образца/пробы (например, отсутствие клеток в биоптате, полученном при пункционной биопсии опухоли) [43, 64, 65]. В некоторых случаях результаты теста не рассматриваются как положительные или отрицательные, как в случае вентиляционно-перфузионного сканирования лёгких при подозрении на лёгочную эмболию, когда результаты классифицируют по трём категориям — нормальные, высоковероятные и неопределённые [66].

Частота сомнительных результатов — важный показатель выполнимости теста, который обычно ограничивает его клиническую ценность, поэтому авторам следует сообщать о таких результатах с указанием причин их возникновения, а также о безуспешном завершении процедуры тестирования. Это касается как индексного, так и референсного тестов.

Игнорирование сомнительных результатов может привести к систематическим ошибкам в оценке точности теста, если только речь идёт не о случайных ошибках. Решение о том, как поступать с такими результатами, может определяться клинической практикой.

Существуют несколько способов обработки сомнительных результатов теста при анализе точности и эффективности теста [63]. Их можно полностью игнорировать, о них можно сообщать, но не учитывать или не рассматривать как отдельную категорию результатов тестирования. Последний вариант особенно полезен, если сомнительные результаты возникают чаще, например, у лиц без целевого состояния, чем у тех, у кого целевое состояние обнаружено. Такие результаты могут классифицировать как ложноположительные или ложноотрицательные в зависимости от результатов референсного теста («наихудший сценарий», *worst-case scenario*) или как истинно положительные и истинно отрицательные («наилучший сценарий», *best-case scenario*).

В примере выше авторы явно выбрали консервативный подход, рассматривая все сомнительные результаты индексного теста как ложноотрицательные (для тех, у кого есть целевое состояние) или ложноположительные (для всех других), — стратегия, которую иногда называют «наихудший сценарий».

Пункт 16. ***Действия в отношении отсутствующих (неполных) данных индексного и референсного тестов***

Пример

«По одной артерии отсутствовали результаты измерения фракционного резерва кровотока, по двум артериям — все данные КТ. Эти артерии исключали из анализа. В качестве альтернативы выполняли замену отсутствующих данных по принципу «наихудшего сценария» (*worst-case imputation*) [67].

Пояснение

Отсутствующие данные — частое явление в любых биомедицинских исследованиях. В диагностических исследованиях такие случаи могут иметь место как для индексного, так и референсного тестов. Есть несколько способов справиться с этой проблемой при анализе данных [68]. Многие исследователи исключают из анализа пациентов без результатов диагностического теста (стратегия анализа «завершённых» или «доступных» случаев). Это может приводить к снижению точности и систематическим ошибкам, особенно если отсутствие результатов индексного или референсного тестов связано с целевым (изучаемым) состоянием.

Участники с отсутствующими результатами теста могут быть включены в анализ, если осуществляется подстановка данных [68–70]. Другой вариант — оценить влияние отсутствующих результатов тестирования на показатели точности с учётом различных сценариев. Для индексного теста, например, «наихудшим сценарием» будет, если все отсутствующие результаты будут считаться ложноположительными или ложноотрицательными в зависимости от результатов референсного теста, а «наилучшим сценарием» — истинно положительными или истинно отрицательными.

В приведённом выше примере авторы сообщили число случаев с отсутствующими данными индексного теста и указали способ их обработки — исключение из анализа согласно «наихудшему» сценарию.

Пункт 17. ***Анализ вариабельности диагностической точности с дифференциацией запланированного и установленного после получения данных результатов разведочного анализа***

Пример

«Чтобы оценить эффективность показателей анализа мочи или их изменение в течение первых 24 часов с целью отличить транзиторное острое повреждение почек от персистирующего, мы построили ROC-кривые для доли истинно положительных против доли ложноположительных результатов, руководствуясь прогностическим правилом для классификации пациентов как больных персистирующим острым повреждением почек. Аналогичную стратегию использовали для оценки эффективности показателей и их изменений во времени в двух предварительно определённых подгруппах пациентов. В первой группе пациенты не получали лечения диуретиками, во второй — не имели признаков сепсиса» [71].

Пояснение

Относительная доля ложноположительных или ложноотрицательных результатов диагностического теста может варьировать в зависимости от характеристик пациента, квалификации исследователей-аналитиков, условий проведения и результатов предыдущих тестов [2, 3]. Как следствие, исследователи могут изучать источники вариабельности точности тестов, оценивая различия полученных результатов между подгруппами пациентов, исследователей или участвующих учреждений.

Апостериорный анализ, который выполняется после просмотра данных, сопряжён с высоким риском ложных результатов. Как правило, такие результаты не подтверждаются последующими исследованиями. Анализ, предварительно описанный в протоколе исследования до сбора данных, вызывает больше доверия [72].

В примере выше авторы сообщили, что точность показателей анализа мочи оценивали в двух заранее отмеченных подгруппах пациентов.

Пункт 18. ***Запланированный размер выборки и его определение***

Пример

«Набор в исследование проводился исходя из предположения, что распространённость аденом размером 6 мм и более в когорте скрининга составит 12%, а точечная оценка чувствительности для этих поражений — 80%. Мы планировали набрать около 600 участников, чтобы добиться предельной погрешности выборки для показателя чувствительности примерно в 8 процентных пунктов. Такой объём выборки также позволял с 90%-ной мощностью обнаружить различия в чувствительности между компьютерной томографической колонографией и оптической колоноскопией, которые могут составлять 18 процентных пунктов и более [73].

Пояснение

Расчёт размера выборки на этапе разработки диагностического исследования может гарантировать достижение достаточной точности. При расчёте размера выборки учитываются конкретные цели исследования и выдвигаемые гипотезы.

Читателям следует сообщать о том, как был определён размер выборки; соответствуют ли сделанные в расчётах допущения научным и клиническим предпосылкам, а также целям исследования; удалось ли авторам набрать запланированное число участников. Методы расчёта размера выборки широко доступны [74–76], но такие расчёты не всегда выполняют или приводят в отчётах диагностических исследований [77, 78].

Размер выборки во многих диагностических исследованиях небольшой. Систематический обзор исследований, опубликованных в 8 ведущих журналах в 2002 г., показал, что медиана размера выборки составляет 118 участников (межквартильный размах от 71 до 350 человек) [77]. Оценки диагностических характеристик тестов в небольших исследованиях, как правило, неточны, с широкими доверительными интервалами.

В приведённом выше примере авторы подробно рассказали о желаемом уровне точности при ожидаемой чувствительности 80%.

Результаты

Пункт 19. ***Формирование выборки исследования***

Пример

«В период с 1 июня 2008 г. по 30 июня 2011 г. оценили 360 пациентов на предмет соответствия первоначальным критериям отбора и пригласили к участию в исследовании. Схема отображает поток пациентов в ходе исследования и первичный исход — развитие прогрессирующей колоректальной неоплазии. Отметили пациентов, исключённых (с указанием причин) или выбывших из исследования. В общей сложности исследование завершили 229 (64%) участников [79] (рис. 1).

Пояснение

Оценки диагностической точности могут быть подвержены систематическим смещениям в том случае, если не все отобранные участники проходят индексный и референсный тесты [80–86] или некоторые из участников проходят другой референсный тест [70]. Неполные данные по референсному тесту наблюдаются в 26% диагностических исследованиях, особенно часто в случаях, где референсный тест — инвазивная процедура [84].

Авторам предлагается приводить в своих отчётах схемы (поточные диаграммы), отображая таким образом последовательность формирования выборки исследования, чтобы читатели могли судить о возможности систематических ошибок. Эта же схема позволит наглядно проиллюстрировать основную структуру исследования. Ниже представлен типичный пример такой схемы (рис. 2).

Представляя точное количество участников для каждого этапа исследования, включая количество истинно положительных, ложноположительных, истинно отрицательных и ложноотрицательных результатов индексного теста, схема исследования также помогает определить правильный знаменатель для расчёта пропорций, таких как чувствительность и специфичность. Помимо этого, диаграмма должна содержать сведения о количестве скринированных участников (*assessed for eligibility*), количестве лиц, которые не прошли индексный и/или референсный тесты с указанием причин. Эта информация поможет читателям оценить риск систематических ошибок, осуществимость стратегии тестирования и применимость результатов исследования.

В примере выше авторы очень кратко описали поток участников и привели схему исследования в виде диаграммы, где отображено количество участников и соответствующие результаты тестирования, полученные

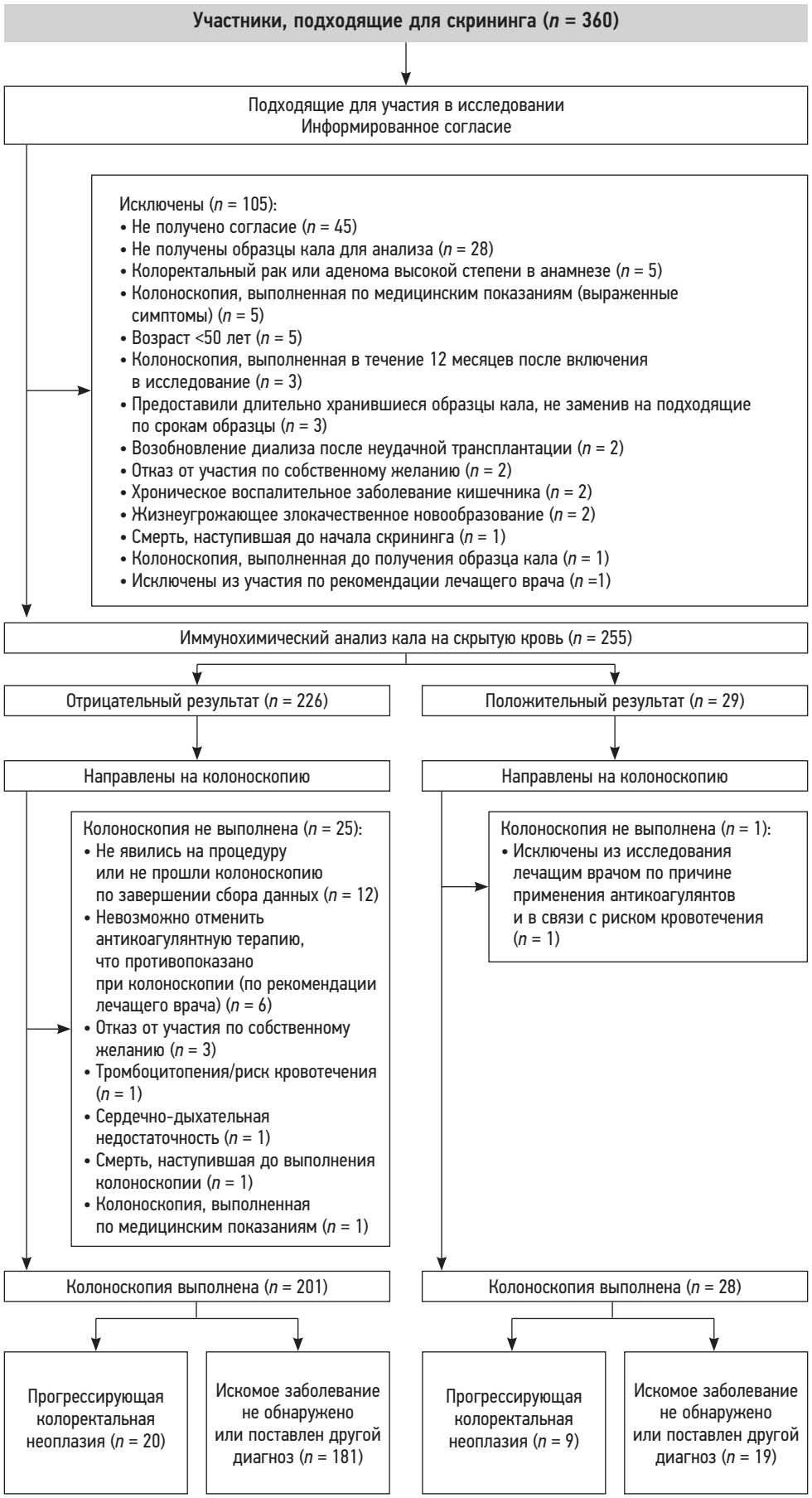

Рис. 1. Пример схемы исследования, в котором оценивалась точность иммунохимического анализа кала для диагностики прогрессирующей колоректальной неоплазии (адаптировано из работы M. Collins и соавт. [79]; публикуется с разрешения).

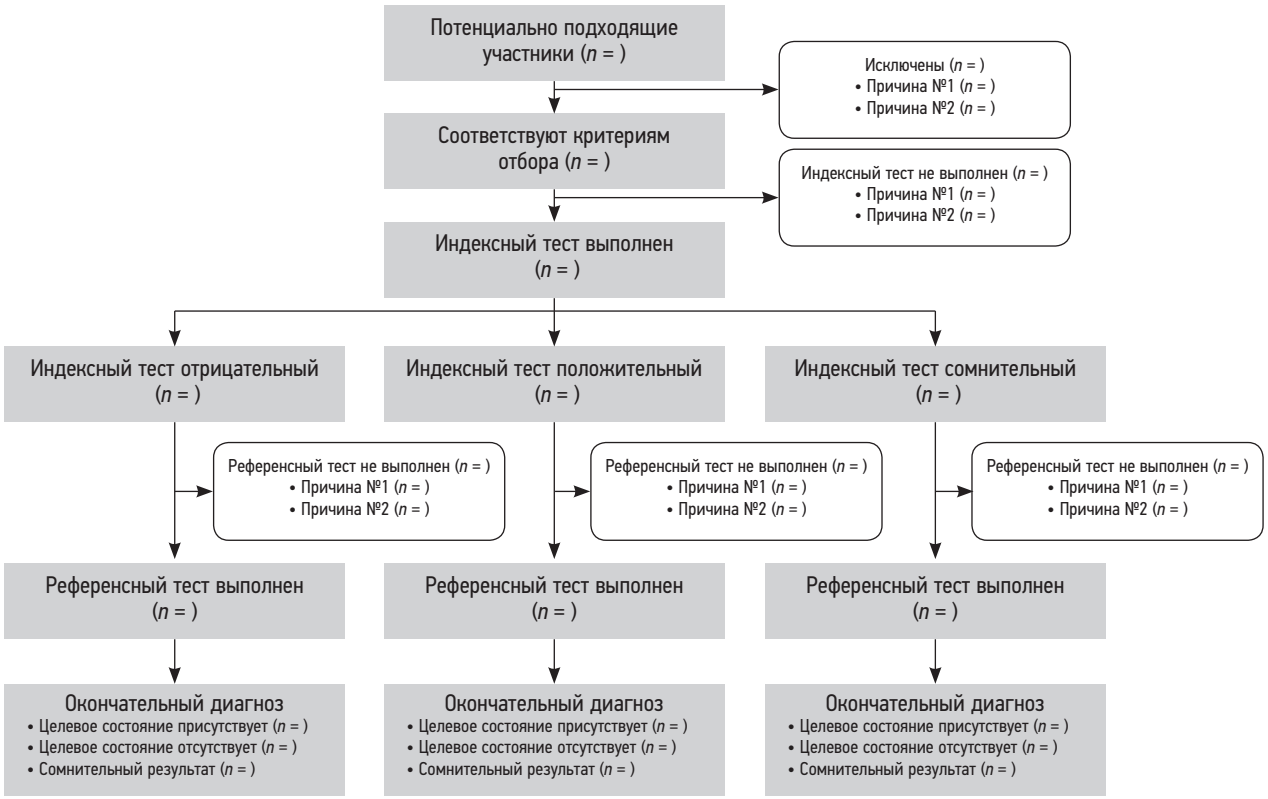

Рис. 2. STARD 2015: поточная диаграмма.

на каждом этапе исследования, с указанием подробных причин исключения участников (см. рис. 1).

Пункт 20. ***Характеристика участников исследования (демографические и клинические данные)***

Пример

«Медианный возраст участников составлял 60 лет (диапазон 18–91), 209 участников (54,7%) были женского пола. Наиболее частые жалобы — боль в животе, затем ректальное кровотечение и диарея, реже отмечались лихорадка и потеря веса. При объективном обследовании пальпация вызывала боль в животе почти у половины пациентов, но пальпируемое новообразование в брюшной полости или прямой кишке обнаружено только у 13 из них (Таблица X)» [87] (табл. 3).

Пояснение

Диагностическая точность теста может зависеть от демографических и клинических характеристик популяции, в которой он применяется [2, 3, 88–92]. Различия по этим характеристикам могут отражать вариабельность степени или тяжести заболевания, что влияет на чувствительность теста, или в альтернативных основному заболеванию состояниях могут приводить к ложноположительным результатам, влияющим на специфичность теста [85].

Адекватное описание демографических и клинических характеристик участников исследования позволяет читателю судить, может ли исследование адекватно ответить на поставленный исследовательский вопрос, и применимы ли результаты исследования к клиническим задачам читателя.

В примере выше авторы представили демографические и клинические данные участников исследования в отдельной таблице. Как правило, это наиболее информативный способ представления основных характеристик участников (см. табл. 3).

Пункт 21a. ***Распределение пациентов с целевым состоянием по тяжести заболевания***

Пункт 21б. ***Распределение пациентов без целевого состояния по альтернативным диагнозам***

Пример

«Из 170 пациентов с ишемической болезнью сердца у 1 пациента обнаружено поражение левой главной коронарной артерии, у 53 — трёхсосудистое, у 64 — двухсосудистое, у 52 — однососудистое поражение коронарного русла. Средняя фракция выброса у пациентов составляла 64% (диапазон 37–83). У остальных 52 пациентов изменений коронарных артерий в результате ангиографии не обнаружено или они были незначительными» [93].

Пояснение

Большинство целевых состояний (объектов диагностического тестирования) не являются чем-то однозначным — присутствующими или отсутствующими. Многие заболевания проходят в своём развитии непрерывный путь от незначительных патологических изменений до клинически выраженных заболеваний. Чувствительность теста

Таблица 3. Пример исходных демографических и клинических характеристик участников в исследовании точности анализа кала с использованием теста Point-of-Care (диагностика по месту лечения) для диагностики органического заболевания кишечника (адаптировано из работы L. Kok и соавт. [87]; публикуется с разрешения)

Характеристики пациентов	*n* (%)
Географический регион проживания в Нидерландах	
• Центральный (Gelderse Vallei)	257 (66,6)
• Южный (Oostelijke Mijnstreek)	129 (33,4)
Медиана возраста (диапазон)	60 (18–91)
Женщины	211 (54,7)
Симптомы при включении	
• Ректальное кровотечение	141 (37,7)
• Боль в животе	267 (70,6)
• Медиана продолжительности боли в животе (диапазон)	150 сут (от 1 сут до 30 лет)
• Стойкая диарея	40 (16,9)
• Диарея	131 (37,2)
• Лихорадка	40 (11,0)
• Снижение веса	62 (17,1)
• Вздутие живота	195 (53,6)
• Запор	169 (46,6)
Физическое обследование	
• Боль при пальпации	117 (46,8)
• Пальпируемое образование в брюшной полости	12 (3,0)
• Пальпируемые фекальные массы	1 (0,3)

часто выше в тех исследованиях, в которых у большего числа пациентов отмечается более тяжёлое течение целевого состояния (*target condition*), поскольку его проще обнаружить посредством индексного теста [28, 85]. Тип, спектр и частота альтернативных диагнозов у пациентов без целевого состояния также могут влиять на точность теста. Как правило, чем лучше себя чувствуют пациенты без целевого состояния, тем реже индексный тест дает ложноположительные результаты [28].

Авторам рекомендуется включать в отчёт информацию о тяжести заболевания у пациентов с целевым состоянием и альтернативных диагнозов у больных без него, что позволит читателям сделать выводы о достоверности (*validity*) исследования относительно поставленного вопроса и применимости результатов исследования к собственным клиническим задачам.

В примере выше авторы исследовали точность тестов с физической нагрузкой для диагностики ишемической болезни сердца. Они сообщили о распределении пациентов по тяжести болезни, выраженной количеством поражённых коронарных артерий (чем их больше, тем более тяжёлую форму имеет заболевание). Чувствительность тестов была выше у пациентов с бóльшим количеством поражённых сосудов: 39% при однососудистом, 58% при двухсосудистом, 77% при трёхсосудистом поражении коронарного русла [91].

Пункт 22. ***Временной интервал и любые медицинские вмешательства между выполнением индексного и референсного тестов***

Пример

«Среднее время между артрометрическим исследованием и магнитно-резонансной томографией составило 38,2 суток (диапазон от 0 до 107 суток)» [94].

Пояснение

Исследования диагностической точности — это, по сути, одномоментные (поперечные, *cross-sectional*) исследования. В большинстве случаев нужно определить, насколько хорошо индексный тест классифицирует пациентов в сравнении с референсным тестом, когда оба теста проводятся у одних и тех же пациентов в одно и то же время [30]. Если один тест выполняется позже другого, целевое и альтернативные состояния пациентов могут изменяться — ухудшаться или улучшаться — вследствие естественного течения заболевания или клинических вмешательств, применяемых в период между двумя тестами. Такие изменения могут влиять на согласованность индексного и референсного тестов, что может привести к систематическим ошибкам оценок эффективности тестов.

Подобные ошибки могут быть более серьёзными, если тесты с положительными и отрицательными результатами или тесты у пациентов с высоким и низким

риском обнаружения целевого состояния систематически проводятся с разными интервалами во времени [1, 2].

Если исследователи рассматривают последующее наблюдение за пациентами в качестве референсного стандарта, необходимо указать продолжительность такого наблюдения.

В приведённом выше примере авторы указали среднее количество дней, а также диапазон значений времени между выполнением индексного и референсного тестов.

Пункт 23. ***Таблицы сопряжённости (или распределение) результатов индексного и референсного тестов***

Пример

«В таблице X **представлены результаты оценки усиления боли в животе при проезде через искусственные неровности в диагностике аппендицита**» [95] (табл. 4).

Пояснение

Результаты исследований должны быть воспроизводимы и доступны для проверки другими учёными. Это относится к процедурам тестирования, проведения исследования и статистическому анализу.

Сопоставление результатов индексного и референсного тестов с помощью таблиц сопряжённости (*cross tabulation*) упрощает пересчёт показателей диагностической точности. Такое представление результатов позволяет определить долю участников с целевым состоянием в выборке исследования, что полезно, поскольку чувствительность и специфичность теста могут варьировать в зависимости от распространённости заболевания [32, 96]. Кросс-табулирование позволяет выполнять также альтернативные или дополнительные анализы (к примеру, метаанализ).

Такие таблицы должны включать абсолютные значения, а не только проценты, поскольку авторы нередко допускают ошибки при вычислении показателей чувствительности и специфичности.

В примере выше авторы представили таблицу сопряжённости, по которой можно легко определить число истинно положительных, ложноположительных, ложноотрицательных и истинно отрицательных результатов (см. табл. 4).

Пункт 24. ***Оценка диагностического показателя и её точность (например, 95% доверительный интервал)***

Пример

«У 46 пациентов в результате компьютерной томографии обнаружен фиброз лёгких. Чувствительность магнитно-резонансной томографии при выявлении этого же диагноза составила 89% (95% ДИ 77–96), специфичность — 91% (95% ДИ 76–98), положительная прогностическая значимость — 93% (95% ДИ 82–99), отрицательная — 86% (95% ДИ 70–95)» [97].

Пояснение

Диагностические исследования никогда не определяют «истинную» чувствительность и специфичность теста. В лучшем случае данные, собранные в ходе исследования, можно использовать для расчёта обоснованных оценок чувствительности и специфичности. Чем меньше участников исследования, тем менее точными (*precise*) будут эти оценки [98].

Наиболее часто используемым выражением неточности является сообщение не только об оценках, иногда называемых точечными оценками (*point estimates*), но и о 95% доверительных интервалах для этих оценок. Результаты исследований с неточными оценками диагностических показателей следует интерпретировать с осторожностью, поскольку за ними скрывается излишний оптимизм авторов [22].

В приведённом примере, где МРТ — индексный тест, а КТ — референсный, авторы указали точечные оценки и 95% доверительные интервалы для показателей чувствительности, специфичности, положительной и отрицательной прогностической значимости (*positive/negative predictive value*).

Пункт 25. ***Любые нежелательные последствия выполнения индексного или референсного тестов***

Пример

«В результате проведения колоноскопии не было отмечено каких-либо серьёзных нежелательных последствий. У 4 пациентов (2%) отмечалось незначительное кровотечение в связи с эндоскопической полипэктомией. Другие незначительные осложнения описаны в приложении» [79].

Таблица 4. Пример таблицы сопряжённости из исследования, в котором изучалась точность использования критерия «усиление боли при проезде через искусственные неровности» в диагностике острого аппендицита (адаптировано из работы H. Ashdown и соавт. [95]; публикуется с разрешения)

	Аппендицит		
Боль в животе при проезде через искусственную неровность	**Положительный тест**	**Отрицательный тест**	**Всего**
Усиливается	33	21	54
Не усиливается	1	9	10
Всего	34	30	64

Пояснение

Не все медицинские тесты одинаково безопасны, и в этом они не отличаются от многих других медицинских вмешательств [99, 100]. Процедура диагностики может привести к различным осложнениям, таким как перфорация при эндоскопии, рентгеноконтрастные реакции аллергического типа при КТ или клаустрофобия при МРТ.

Измерение и регистрация нежелательных явлений в диагностических исследованиях предоставят дополнительную информацию клиницистам, которые могут неохотно использовать их, если они вызывают серьёзные или частые нежелательные явления. Фактическое применение индексного теста в клинической практике определяется не только его точностью, но и другими аспектами, включая его выполнимость и безопасность. Это также относится и к референсному тесту.

В представленном примере авторы различают «значительные» и «незначительные» нежелательные последствия проведения диагностического исследования и сообщают, как часто они возникают.

Обсуждение

Пункт 26. ***Ограничения исследования, включая источники потенциальных систематических ошибок, статистической неопределённости и ограниченной обобщаемости результатов***

Пример

«Исследование было сопряжено с рядом ограничений. Во-первых, не всех пациентов, которые прошли процедуру КТ-колонографии (КТК), оценивали референсными методами. <...> Мы исключили из исследования 41 пациента, которые соответствовали критериям отбора, но не прошли референсные процедуры и имели отрицательные или умеренно положительные результаты КТК, что могло привести к слегка завышенным показателям чувствительности КТК (т.е. имела место систематическая ошибка верификации, *partial verification bias*). Во-вторых, в некоторых случаях (преимущественно у пациентов с отрицательными результатами) были большие интервалы времени между проведением КТК и референсного метода диагностики. <...> Во всяком случае, увеличенный интервал, предположительно, немного занижает чувствительность и отрицательную прогностическую значимость КТК при диагностике доброкачественных образований, поскольку «пропущенные» образования могли гипотетически развиться или увеличиться в размере с момента выполнения КТК» [101].

Пояснение

Диагностические исследования подвержены риску систематических ошибок, как и другие клинические испытания и исследования. В результате авторы могут получать оценки точности, которые не отражают действительные характеристики теста в связи с ошибками и недостатками дизайна исследования или анализа данных [1, 2]. Вследствие различий в дизайне, участниках и процедурах результаты одного конкретного диагностического исследования могут оказаться невоспроизводимыми в других условиях, а их обобщаемость (*generalisability*) будет носить ограниченный характер [102].

В разделе «Обсуждение» авторы должны критически оценивать достоверность (*validity*) полученных данных, отмечать потенциальные ограничения и уточнять, по какой причине можно или нельзя распространить полученные результаты на другие условия. Поскольку систематические ошибки могут сводиться к переоценке или недооценке точности индексного теста, авторам следует обсудить направление возможного смещения вместе с его вероятной величиной. Затем читателей необходимо проинформировать о вероятности того, что ограничения исследования ставят под угрозу его результаты и выводы (см. пункт 27) [103].

Некоторые журналы прямо призывают авторов сообщать об ограничениях исследования, но многие не конкретизируют, какие элементы должны быть рассмотрены [104]. Для диагностических исследований мы настоятельно рекомендуем обсудить как минимум возможные источники систематических ошибок, неточность данных и вопросы, связанные с набором пациентов и условиями, в которых проводилось исследование.

В примере выше авторы определили два возможных источника систематических ошибок, характерных для диагностических исследований: (1) не все результаты теста верифицировали референсным методом диагностики (*partial verification bias*) и (2) между выполнением индексного и референсного тестов был временной интервал, в течение которого целевое состояние могло измениться. Авторы также обсудили величину потенциальных систематических ошибок и их направление, уточнив, могли ли они привести к переоценке или недооценке точности теста.

Пункт 27. ***Значение для практики, включая предполагаемое использование и клиническую роль индексного теста***

Пример

«Оценка по шкале Уэллса <4 баллов в сочетании с отрицательным тестом на D-димер позволили исключить развитие лёгочной тромбоэмболии у 4–5 пациентов из 10 с частотой ошибок менее 2%, что считается безопасным, согласно большинству опубликованных рекомендаций. Такая стратегия позволяет врачам первичного звена здравоохранения безопасно исключать лёгочную тромбоэмболию у значительного числа пациентов с подозрением на наличие такого заболевания, тем самым уменьшая затраты и нагрузку на пациента (например, снижая риск развития контрастиндуцированной нефропатии, ассоциированной с мультиспиральной компьютерной томографией), связанные с ненужным

направлением в учреждения специализированной медицинской помощи» [25].

Пояснение

Для того чтобы результаты исследования были актуальными для практики, авторам диагностических исследований следует подробно описать последствия своих выводов, принимая во внимание предполагаемое использование (цель тестирования) и клиническую роль теста (место теста в существующих клинических протоколах ведения больных).

Тест может быть предложен для диагностики, определения предрасположенности, скрининга, стратификации риска, определения стадии заболевания, предсказания (*prediction*), прогнозирования (*prognosis*), выбора лечения, мониторинга, надзора или других целей. Клиническая роль теста (предварительный тест, дополнительный тест, замещающий тест) отражает его положение по отношению к другим тестам, выполняемым с аналогичной целью и в аналогичных условиях [19, 105]. Предполагаемое использование и клиническая роль индексного теста должны быть описаны в вводном разделе статьи (см. Пункт 3).

Предполагаемое использование и роль теста будут определять желаемую величину показателей диагностической точности. Например, для исключения заболевания с помощью недорогого предварительного теста (*triage test*) требуется высокая чувствительность, и вместе с тем допустима неидеальная специфичность. Если же тест предназначен для подтверждения заболевания, специфичность может стать гораздо более важной характеристикой теста [106].

В разделе «Обсуждение» авторы должны уточнить, соответствуют или нет полученные оценки точности теста целям исследования.

В приведённом выше примере авторы пришли к выводу, что оценка вероятности развития лёгочной тромбоэмболии, составляющая <4 баллов по шкале Уэллса, в сочетании отрицательным результатом теста на D-димер, выполненного у постели больного, позволяли исключить лёгочную тромбоэмболию у большинства пациентов, которые обращались за первичной медицинской помощью.

Дополнительная информация

Пункт 28. ***Регистрационный номер исследования и наименование регистра***

Пример

«Исследование зарегистрировано на http://www.clinicaltrials.org (NCT00916864)» [107].

Пояснение

Регистрация протоколов исследований до их начала в регистре клинических испытаний, например на ClinicalTrials.gov или в одном из первичных регистров Всемирной организации здравоохранения, позволяет без труда идентифицировать в базе данных то или иное исследование [108–112]. Это даёт много преимуществ, в том числе позволяет избежать необоснованного частичного или полного повторения исследований, а также позволит коллегам и потенциальным участникам связаться с координаторами исследования.

Дополнительные преимущества регистрации исследований — проспективное определение целей исследования, показателей исхода, критериев отбора и данных, которые необходимо собрать, что позволит редакторам, рецензентам и читателям идентифицировать отклонения в финальном отчёте исследования. Регистрация испытаний, кроме того, позволяет идентифицировать исследования, которые были завершены, но ещё не опубликованы.

Многие журналы требуют регистрации клинических испытаний. Доля регистрируемых диагностических исследований непрерывно растёт, хотя и остаётся небольшой. По результатам недавней оценки, из 351 диагностического исследования, опубликованного в журналах с высоким импакт-фактором в 2012 году, только 15% были предварительно зарегистрированы [113].

Включение регистрационного номера в отчёт об исследовании облегчает его поиск в соответствующем регистре. Более того, регистрацию исследования до начала его проведения можно считать признаком его качества.

В представленном выше примере авторы сообщили, что исследование было зарегистрировано в регистре ClinicalTrials.gov с указанием регистрационного номера, чтобы можно было легко найти соответствующую запись.

Пункт 29. ***Доступ к полному протоколу исследования***

Пример

«Более подробная информация о дизайне и обосновании исследования OPTIMAP была опубликована ранее [ссылка на протокол исследования]» [114].

Пояснение

Полные протоколы исследования обычно включают дополнительную методологическую информацию, которая не представлена в окончательном отчёте из-за ограничений по количеству публикуемых слов или по причине того, что эти данные уже были опубликованы в других источниках. Такая информация может быть полезна для тех, кто хочет в полной мере оценить достоверность исследования, воспроизвести или применить на практике процедуры тестирования.

Всё большее число авторов публикуют первоначальный протокол исследования, часто до момента включения в исследование первого участника. Протоколы зачастую публикуют в научных журналах, на веб-сайтах медицинских учреждений или спонсоров, или в качестве дополнительных материалов на веб-сайте журнала, в котором будет опубликован отчёт об исследовании.

Если протокол опубликован или размещён онлайн, авторы должны предоставить соответствующую

библиографическую ссылку или ссылку на электронный документ. Если протокол исследования не был опубликован, авторы должны указать, у кого его можно получить [115].

В примере выше авторы привели библиографическую ссылку на полную версию протокола, который был опубликован ранее.

Пункт 30. ***Источники финансирования, другие виды поддержки и роль спонсоров исследования***

Пример

«Финансирование приобретения дополнительных диагностических реагентов и оборудования, необходимых для исследования, предоставлено компанией Gen-Probe. Спонсоры не принимали участия в инициировании или разработке исследования, сборе образцов, анализе и интерпретации данных, написании статьи и её представлении к печати. Исследование и исследователи были независимыми от спонсоров, компании Gen-Probe» [116].

Пояснение

Известно, что спонсирование исследования фармацевтической компанией связано с получением результатов, благоприятствующих интересам этого спонсора [117]. К сожалению, информация о спонсорах зачастую не раскрывается в научных статьях, что затрудняет оценку связанных с этим потенциальных систематических ошибок. Спонсорство может заключаться в прямом финансировании исследования или в предоставлении основных материалов для его проведения, в том числе устройства для тестирования.

Роль спонсоров, включая степень их участия в исследовании, варьирует. Спонсор может, например, участвовать в разработке и проведении исследования, а также анализе данных, составлении отчётов и принятии решения о публикации. Авторам рекомендуется чётко указывать источники финансирования, а также роль спонсоров в исследовании, поскольку такая прозрачность помогает читателям оценить уровень независимости исследователей.

В приведённом выше примере авторы сообщили о том, в какой мере была оказана спонсорская поддержка, а также о своей независимости на каждом этапе исследования.

ДОПОЛНИТЕЛЬНО

Вклад авторов. J.F. Cohen, D.A. Korevaar внесли равный вклад в написание этой статьи и в одинаковой мере являются первыми авторами; J.F. Cohen, D.A. Korevaar, P.M.M. Bossuyt — написание черновика рукописи; D.G. Altman, D.E. Bruns, C.A. Gatsonis, L. Hooft, L. Irwig, D. Levine, J.B. Reitsma, H.C.W. de Vet — критический пересмотр и редактирование рукописи.

Благодарности. Авторы благодарят инициативную группу STARD за помощь в определении ключевых пунктов для отчётов о диагностических исследованиях.

Заявление о доступе к данным. Дополнительные данные не предоставлены.

СПИСОК ЛИТЕРАТУРЫ

1. Whiting P.F., Rutjes A.W., Reitsma J.B., et al. Sources of variation and bias in studies of diagnostic accuracy: a systematic review // Ann Intern Med. 2004. Vol. 140, N 3. P. 189–202. doi: 10.7326/0003-4819-140-3-200402030-00010

2. Whiting P.F., Rutjes A.W., Westwood M.E., et al. A systematic review classifies sources of bias and variation in diagnostic test accuracy studies // J Clin Epidemiol. 2013. Vol. 66, N 10. P. 1093–1104. doi: 10.1016/j.jclinepi.2013.05.014

3. Whiting P.F., Rutjes A.W., Westwood M.E., et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies // Ann Intern Med. 2011. Vol. 155, N 8. P. 529–536. doi: 10.7326/0003-4819-155-8-201110180-00009

4. Korevaar D.A., van Enst W.A., Spijker R., et al. Reporting quality of diagnostic accuracy studies: a systematic review and meta-analysis of investigations on adherence to STARD // Evid Based Med. 2014. Vol. 19, N 2. P. 47–54. doi: 10.1136/eb-2013-101637

5. Korevaar D.A., Wang J., van Enst W.A., et al. Reporting diagnostic accuracy studies: some improvements after 10 years of STARD // Radiology. 2015. Vol. 274, N 3. P. 781–789. doi: 10.1148/radiol.14141160

6. Lijmer J.G., Mol B.W., Heisterkamp S., et al. Empirical evidence of design-related bias in studies of diagnostic tests // JAMA. 1999. Vol. 282, N 11. P. 1061–1066. doi: 10.1001/jama.282.11.1061

7. Bossuyt P.M., Reitsma J.B., Bruns D.E., et al. Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative // Clin Chem. 2003. Vol. 49, N 1. P. 1–6. doi: 10.1373/49.1.1

8. Begg C., Cho M., Eastwood S., et al. Improving the quality of reporting of randomized controlled trials. The CONSORT statement // JAMA. 1996. Vol. 276, N 8. P. 637–639. doi: 10.1001/jama.276.8.637

9. Schulz K.F., Altman D.G., Moher D. CONSORT 2010 statement: updated guidelines for reporting parallel group randomised trials // BMJ. 2010. Vol. 340, N 1. P. 332. doi: 10.1136/bmj.c332

10. Bossuyt P.M., Reitsma J.B., Bruns D.E., et al. STARD 2015: an updated list of essential items for reporting diagnostic accuracy studies // BMJ. 2015. Vol. 351. P. h5527. doi: 10.1136/bmj.h5527

11. Bossuyt P.M., Reitsma J.B., Bruns D.E., et al. The STARD statement for reporting studies of diagnostic accuracy: explanation and elaboration // Ann Intern Med. 2003. Vol. 138, N 1. P. W1–12. doi: 10.7326/0003-4819-138-1-200301070-00012-w1

12. Regge D., Laudi C., Galatola G., et al. Diagnostic accuracy of computed tomographic colonography for the detection of advanced neoplasia in individuals at increased risk of colorectal cancer // JAMA. 2009. Vol. 301, N 23. P. 2453–2461. doi: 10.1001/jama.2009.832

13. Deville W.L., Bezemer P.D., Bouter L.M. Publications on diagnostic test evaluation in family medicine journals: an optimal search strategy // J Clin Epidemiol. 2000. Vol. 53, N 1. P. 65–69. doi: 10.1016/s0895-4356(99)00144-4

14. Korevaar D.A., Cohen J.F., Hooft L., et al. Literature survey of high-impact journals revealed reporting weaknesses in abstracts of diagnostic accuracy studies // J Clin Epidemiol. 2015. Vol. 68, N 6. P. 708–715. doi: 10.1016/j.jclinepi.2015.01.014
15. Korevaar D.A., Cohen J.F., de Ronde M.W., et al. Reporting weaknessess in conference abstracts of diagnostic accuracy studies in ophthalmology // JAMA Ophthalmol. 2015. Vol. 133, N 12. P. 1464–1467. doi: 10.1001/jamaophthalmol.2015.3577
16. A proposal for more informative abstracts of clinical articles. Ad Hoc Working Group for Critical Appraisal of the Medical Literature // Ann Intern Med. 1987. Vol. 106, N 4. P. 598–604.
17. Stiell I.G., Greenberg G.H., Wells G.A., et al. Derivation of a decision rule for the use of radiography in acute knee injuries // Ann Emerg Med. 1995. Vol. 26, N 4. P. 405–413. doi: 10.1016/s0196-0644(95)70106-0
18. Horvath A.R., Lord S.J., StJohn A., et al. From biomarkers to medical tests: the changing landscape of test evaluation // Clin Chim Acta. 2014. Vol. 427. P. 49–57. doi: 10.1016/j.cca.2013.09.018
19. Bossuyt P.M., Irwig L., Craig J., et al. Comparative accuracy: assessing new tests against existing diagnostic pathways // BMJ. 2006. Vol. 332. P. 1089–1092. doi: 10.1136/bmj.332.7549.1089
20. Gieseker K.E., Roe M.H., MacKenzie T., et al. Evaluating the American Academy of Pediatrics diagnostic standard for Streptococcus pyogenes pharyngitis: backup culture versus repeat rapid antigen testing // Pediatrics. 2003. Vol. 111, N 6 Pt 1. P. e666–670. doi: 10.1542/peds.111.6.e666
21. Tanz R.R., Gerber M.A., Kabat W., et al. Performance of a rapid antigen-detection test and throat culture in community pediatric offices: implications for management of pharyngitis // Pediatrics. 2009. Vol. 123, N 2. P. 437–444. doi: 10.1542/peds.2008-0488
22. Ochodo E.A., de Haan M.C., Reitsma J.B., et al. Overinterpretation and misreporting of diagnostic accuracy studies: evidence of 'spin' // Radiology. 2013. Vol. 267, N 2. P. 581–588. doi: 10.1148/radiol.12120527
23. Freer P.E., Niell B., Rafferty E.A. Preoperative tomosynthesis-guided needle localization of mammographically and sonographically occult breast lesions // Radiology. 2015. Vol. 275, N 2. P. 377–383. doi: 10.1148/radiol.14140515
24. Sorensen H.T., Sabroe S., Olsen J. A framework for evaluation of secondary data sources for epidemiological research // Int J Epidemiol. 1996. Vol. 25, N 2. P. 435–442. doi: 10.1093/ije/25.2.435
25. Geersing G.J., Erkens P.M., Lucassen W.A., et al. Safe exclusion of pulmonary embolism using the Wells rule and qualitative D-dimer testing in primary care: prospective cohort study // BMJ. 2012. Vol. 345. P. e6564. doi: 10.1136/bmj.e6564
26. Bomers M.K., van Agtmael M.A., Luik H., et al. Using a dog's superior olfactory sensitivity to identify Clostridium difficile in stools and patients: proof of principle study // BMJ. 2012. Vol. 345. P. e7396. doi: 10.1136/bmj.e7396
27. Philbrick J.T., Horwitz R.I., Feinstein A.R. Methodologic problems of exercise testing for coronary artery disease: groups, analysis and bias // Am J Cardiol. 1980. Vol. 46, N 5. P. 807–812. doi: 10.1016/0002-9149(80)90432-4
28. Rutjes A.W., Reitsma J.B., Vandenbroucke J.P., et al. Case-control and two-gate designs in diagnostic accuracy studies // Clin Chem. 2005. Vol. 51, N 8. P. 1335–1341. doi: 10.1373/clinchem.2005.048595
29. Rutjes A.W., Reitsma J.B., Di Nisio M., et al. Evidence of bias and variation in diagnostic accuracy studies // CMAJ. 2006. Vol. 174, N 4. P. 469–476. doi: 10.1503/cmaj.050090
30. Knottnerus J.A., Muris J.W. Assessment of the accuracy of diagnostic tests: the cross-sectional study // J Clin Epidemiol. 2003. Vol. 56, N 11. P. 1118–1128. doi: 10.1016/s0895-4356(03)00206-3
31. Van der Schouw Y.T., Van Dijk R., Verbeek A.L. Problems in selecting the adequate patient population from existing data files for assessment studies of new diagnostic tests // J Clin Epidemiol.1995. Vol. 48, N 3. P. 417–422. doi: 10.1016/0895-4356(94)00144-f
32. Leeflang M.M., Bossuyt P.M., Irwig L. Diagnostic test accuracy may vary with prevalence: implications for evidence-based diagnosis // J Clin Epidemiol. 2009. Vol. 62, N 1. P. 5–12. doi: 10.1016/j.jclinepi.2008.04.007
33. Attia M., Zaoutis T., Eppes S., et al. Multivariate predictive models for group A beta-hemolytic streptococcal pharyngitis in children // Acad Emerg Med. 1999. Vol. 6, N 1. P. 8–13. doi: 10.1111/j.1553-2712.1999.tb00087.x
34. Knottnerus J.A., Knipschild P.G., Sturmans F. Symptoms and selection bias: the influence of selection towards specialist care on the relationship between symptoms and diagnoses // Theor Med. 1989. Vol. 10, N 1. P. 67–81. doi: 10.1007/BF00625761
35. Knottnerus J.A., Leffers P. The influence of referral patterns on the characteristics of diagnostic tests // J Clin Epidemiol. 1992. Vol. 45, N 10. P. 1143–1154. doi: 10.1016/0895-4356(92)90155-g
36. Melbye H., Straume B. The spectrum of patients strongly influences the usefulness of diagnostic tests for pneumonia // Scand J Prim Health Care. 1993. Vol. 11, N 4. P. 241–246. doi: 10.3109/02813439308994838
37. Ezike E.N., Rongkavilit C., Fairfax M.R., et al. Effect of using 2 throat swabs vs 1 throat swab on detection of group A streptococcus by a rapid antigen detection test // Arch Pediatr Adolesc Med. 2005. Vol. 159, N 5. P. 486–490. doi: 10.1001/archpedi.159.5.486
38. Rosjo H., Kravdal G., Hoiseth A.D., et al. Troponin I measured by a high-sensitivity assay in patients with suspected reversible myocardial ischemia: data from the Akershus Cardiac Examination (ACE) 1 study // Clin Chem. 2012. Vol. 58, N 11. P. 1565–1573. doi: 10.1373/clinchem.2012.190868
39. Irwig L., Bossuyt P., Glasziou P., et al. Designing studies to ensure that estimates of test accuracy are transferable // BMJ. 2002. Vol. 324, N 7338. P. 669–671. doi: 10.1136/bmj.324.7338.669
40. Detrano R., Gianrossi R., Froelicher V. The diagnostic accuracy of the exercise electrocardiogram: a meta-analysis of 22 years of research // Prog Cardiovasc Dis. 1989. Vol. 32, N 3. P. 173–206. doi: 10.1016/0033-0620(89)90025-x
41. Brealey S., Scally A.J. Bias in plain film reading performance studies // Br J Radiol. 2001. Vol. 74, N 880. P. 307–316. doi: 10.1259/bjr.74.880.740307
42. Elmore J.G., Wells C.K., Lee C.H., et al. Variability in radiologists' interpretations of mammograms // N Engl J Med. 1994. Vol. 331, N 22. P. 1493–1499. doi: 10.1056/NEJM199412013312206
43. Ronco G., Montanari G., Aimone V., et al. Estimating the sensitivity of cervical cytology: errors of interpretation and test limitations // Cytopathology. 1996. Vol. 7, N 3. P. 151–158. doi: 10.1046/j.1365-2303.1996.39382393.x
44. Cohen M.B., Rodgers R.P., Hales M.S., et al. Influence of training and experience in fine-needle aspiration biopsy of breast. Receiver

operating characteristics curve analysis // Arch Pathol Lab Med. 1987. Vol. 111, N 6. P. 518–520.
45. Fox J.W., Cohen D.M., Marcon M.J., et al. Performance of rapid streptococcal antigen testing varies by personnel // J Clin Microbiol. 2006. Vol. 44, N 11. P. 3918–3922. doi: 10.1128/JCM.01399-06
46. Gandy M., Sharpe L., Perry K.N., et al. Assessing the efficacy of 2 screening measures for depression in people with epilepsy // Neurology. 2012. Vol. 79, N 4. P. 371–375. doi: 10.1212/WNL.0b013e318260cbfc
47. Stegeman I., de Wijkerslooth T.R., Stoop E.M., et al. Combining risk factors with faecal immunochemical test outcome for selecting CRC screenees for colonoscopy // Gut. 2014. Vol. 63, N 3. P. 466–471. doi: 10.1136/gutjnl-2013-305013
48. Leeflang M.M., Moons K.G., Reitsma J.B., et al. Bias in sensitivity and specificity caused by data-driven selection of optimal cutoff values: mechanisms, magnitude, and solutions // Clin Chem. 2008. Vol. 54, N 4. P. 729–737. doi: 10.1373/clinchem.2007.096032
49. Ewald B. Post hoc choice of cut points introduced bias to diagnostic research // J Clin Epidemiol. 2006. Vol. 59, N 8. P. 798–801. doi: 10.1016/j.jclinepi.2005.11.025
50. Justice A.C., Covinsky K.E., Berlin J.A. Assessing the generalizability of prognostic information // Ann Intern Med. 1999. Vol. 130, N 6. P. 515–524. doi: 10.7326/0003-4819-130-6-199903160-00016
51. Harrell F.E., Lee K.L., Mark D.B. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors // Stat Med. 1996. Vol. 15, N 4. P. 361–387. doi: 10.1002/(SICI)1097-0258(19960229)15:4<361::AID-SIM168>3.0.CO.;2-4
52. Hodgdon T., McInnes M.D., Schieda N., et al. Can quantitative CT texture analysis be used to differentiate fat-poor renal angiomyolipoma from renal cell carcinoma on unenhanced CT images? // Radiology. 2015. Vol. 276, N 3. P. 787–796. doi: 10.1148/radiol.2015142215
53. Begg C.B. Biases in the assessment of diagnostic tests // Stat Med. 1987. Vol. 6, N 4. P. 411–423. doi: 10.1002/sim.4780060402
54. Doubilet P., Herman P.G. Interpretation of radiographs: effect of clinical history // AJR Am J Roentgenol. 1981. Vol. 137, N 5. P. 1055–1058. doi: 10.2214/ajr.137.5.1055
55. D'Orsi C.J., Getty D.J., Pickett R.M., et al. Stereoscopic digital mammography: improved specificity and reduced rate of recall in a prospective clinical trial // Radiology. 2013. Vol. 266, N 1. P. 81–88. doi: 10.1148/radiol.12120382
56. Knottnerus J.A., Buntinx F. The evidence base of clinical diagnosis: theory and methods of diagnostic research. 2nd edn. BMJ Books, 2008. 316 p.
57. Pepe M. Study design and hypothesis testing. The statistical evaluation of medical tests for classification and prediction. Oxford, UK: Oxford University Press, 2003. P. 214–251.
58. Hayen A., Macaskill P., Irwig L., et al. Appropriate statistical methods are required to assess diagnostic tests for replacement, add-on, and triage // J Clin Epidemiol. 2010. Vol. 63, N 8. P. 883–891. doi: 10.1016/j.jclinepi.2009.08.024
59. Pena B.M., Mandl K.D., Kraus S.J., et al. Ultrasonography and limited computed tomography in the diagnosis and management of appendicitis in children // JAMA. 1999. Vol. 282, N 11. P. 1041–1046. doi: 10.1001/jama.282.11.1041
60. Simel D.L., Feussner J.R., DeLong E.R., et al. Intermediate, indeterminate, and uninterpretable diagnostic test results // Med Decis Making. 1987. Vol. 7, N 2. P. 107–114. doi: 10.1177/0272989X8700700208
61. Philbrick J.T., Horwitz R.I., Feinstein A.R., et al. The limited spectrum of patients studied in exercise test research. Analyzing the tip of the iceberg // JAMA. 1982. Vol. 248, N 19. P. 2467–2470.
62. Begg C.B., Greenes R.A., Iglewicz B. The influence of uninterpretability on the assessment of diagnostic tests // J Chronic Dis. 1986. Vol. 39, N 8. P. 575–584. doi: 10.1016/0021-9681(86)90182-7
63. Shinkins B., Thompson M., Mallett S., et al. Diagnostic accuracy studies: how to report and analyse inconclusive test results // BMJ. 2013. Vol. 346. P. f2778. doi: 10.1136/bmj.f2778
64. Pisano E.D., Fajardo L.L., Tsimikas J., et al. Rate of insufficient samples for fine-needle aspiration for nonpalpable breast lesions in a multicenter clinical trial: the Radiologic Diagnostic Oncology Group 5 Study. The RDOG5 investigators // Cancer. 1998. Vol. 82, N 4. P. 679–688. doi: 10.1002/(sici)1097-0142(19980215)82:4<679::aid-cncr10>3.0.co;2-v
65. Giard R.W., Hermans J. The value of aspiration cytologic examination of the breast. A statistical review of the medical literature // Cancer. 1992. Vol. 69, N 8. P. 2104–2110. doi: 10.1002/1097-0142(19920415)69:8<2104::aid-cncr2820690816>3.0.co;2-o
66. Investigators P. Value of the ventilation/perfusion scan in acute pulmonary embolism. Results of the prospective investigation of pulmonary embolism diagnosis (PIOPED) // JAMA. 1990. Vol. 263, N 20. P. 2753–2759. doi: 10.1001/jama.1990.03440200057023
67. Min J.K., Leipsic J., Pencina M.J., et al. Diagnostic accuracy of fractional flow reserve from anatomic CT angiography // JAMA. 2012. Vol. 308, N 12. P. 1237–1245. doi: 10.1001/2012.jama.11274
68. Naaktgeboren C.A., de Groot J.A., Rutjes A.W., et al. Anticipating missing reference standard data when planning diagnostic accuracy studies // BMJ. 2016. Vol. 352. P. i402. doi: 10.1136/bmj.i402
69. Van der Heijden G.J., Donders A.R., Stijnen T., et al. Imputation of missing values is superior to complete case analysis and the missing-indicator method in multivariable diagnostic research: a clinical example // J Clin Epidemiol. 2006. Vol. 59, N 10. P. 1102–1109. doi: 10.1016/j.jclinepi.2006.01.015
70. de Groot J.A., Bossuyt P.M., Reitsma J.B., et al. Verification problems in diagnostic accuracy studies: consequences and solutions // BMJ. 2011. Vol. 343. P. d4770. doi: 10.1136/bmj.d4770
71. Pons B., Lautrette A., Oziel J., et al. Diagnostic accuracy of early urinary index changes in differentiating transient from persistent acute kidney injury in critically ill patients: multicenter cohort study // Crit Care. 2013. Vol. 17, N 2. P. R56. doi: 10.1186/cc12582
72. Sun X., Ioannidis J.P., Agoritsas T., et al. How to use a subgroup analysis: users' guide to the medical literature // JAMA 2014. Vol. 311, N 4. P. 405–411. doi: 10.1001/jama.2013.285063
73. Zalis M.E., Blake M.A., Cai W., et al. Diagnostic accuracy of laxative-free computed tomographic colonography for detection of adenomatous polyps in asymptomatic adults: a prospective evaluation // Ann Intern Med. 2012. Vol. 156, N 10. P. 692–702. doi: 10.7326/0003-4819-156-10-201205150-00005
74. Flahault A., Cadilhac M., Thomas G. Sample size calculation should be performed for design accuracy in diagnostic test studies // J Clin Epidemiol. 2005. Vol. 58, N 8. P. 859–862. doi: 10.1016/j.jclinepi.2004.12.009
75. Pepe M.S. The statistical evaluation of medical tests for classification and prediction. Oxford, New York: Oxford University Press, 2003.

76. Vach W., Gerke O., Hoilund-Carlsen P.F. Three principles to define the success of a diagnostic study could be identified // J Clin Epidemiol. 2012. Vol. 65, N 3. P. 293–300. doi: 10.1016/j.jclinepi.2011.07.004
77. Bachmann L.M., Puhan M.A., ter Riet G., et al. Sample sizes of studies on diagnostic accuracy: literature survey // BMJ. 2006. Vol. 332, N 4550. P. 1127–1129. doi: 10.1136/bmj.38793.637789.2F
78. Bochmann F., Johnson Z., Azuara-Blanco A. Sample size in studies on diagnostic accuracy in ophthalmology: a literature survey // Br J Ophthalmol. 2007. Vol. 91, N 7. P. 898–900. doi: 10.1136/bjo.2006.113290
79. Collins M.G., Teo E., Cole S.R., et al. Screening for colorectal cancer and advanced colorectal neoplasia in kidney transplant recipients: cross sectional prevalence and diagnostic accuracy study of faecal immunochemical testing for haemoglobin and colonoscopy // BMJ. 2012. Vol. 345. P. e4657. doi: 10.1136/bmj.e4657
80. Cecil M.P., Kosinski A.S., Jones M.T., et al. The importance of work-up (verification) bias correction in assessing the accuracy of SPECT thallium-201 testing for the diagnosis of coronary artery disease // J Clin Epidemiol. 1996. Vol. 49, N 7. P. 735–742. doi: 10.1016/0895-4356(96)00014-5
81. Choi B.C. Sensitivity and specificity of a single diagnostic test in the presence of work-up bias // J Clin Epidemiol. 1992. Vol. 45, N 6. P. 581–586. doi: 10.1016/0895-4356(92)90129-b
82. Diamond G.A. Off Bayes: effect of verification bias on posterior probabilities calculated using Bayes' theorem // Med Decis Making. 1992. Vol. 12, N 1. P. 22–31. doi: 10.1177/0272989X9201200105
83. Diamond G.A., Rozanski A., Forrester J.S., et al. A model for assessing the sensitivity and specificity of tests subject to selection bias. Application to exercise radionuclide ventriculography for diagnosis of coronary artery disease // J Chronic Dis. 1986. Vol. 39, N 5. P. 343–355. doi: 10.1016/0021-9681(86)90119-0
84. Greenes R.A., Begg C.B. Assessment of diagnostic technologies. Methodology for unbiased estimation from samples of selectively verified patients // Invest Radiol. 1985. Vol. 20, N 7. P. 751–756.
85. Ransohoff D.F., Feinstein A.R. Problems of spectrum and bias in evaluating the efficacy of diagnostic tests // N Engl J Med. 1978. Vol. 299, N 17. P. 926–930. doi: 10.1056/NEJM197810262991705
86. Zhou X.H. Effect of verification bias on positive and negative predictive values // Stat Med. 1994. Vol. 13, N 17. P. 1737–1745. doi: 10.1002/sim.4780131705
87. Kok L., Elias S.G., Witteman B.J., et al. Diagnostic accuracy of point-of-care fecal calprotectin and immunochemical occult blood tests for diagnosis of organic bowel disease in primary care: the Cost-Effectiveness of a Decision Rule for Abdominal Complaints in Primary Care (CEDAR) study // Clin Chem. 2012. Vol. 58, N 6. P. 989–998. doi: 10.1373/clinchem.2011.177980
88. Harris J.M. The hazards of bedside Bayes // JAMA. 1981. Vol. 246, N 22. P. 2602–2605.
89. Hlatky M.A., Pryor D.B., Harrell F.E., et al. Factors affecting sensitivity and specificity of exercise electrocardiography. Multivariable analysis // Am J Med. 1984. Vol. 77, N 1. P. 64–71. doi: 10.1016/0002-9343(84)90437-6
90. Lachs M.S., Nachamkin I., Edelstein P.H, et al. Spectrum bias in the evaluation of diagnostic tests: lessons from the rapid dipstick test for urinary tract infection // Ann Intern Med. 1992. Vol. 117, N 2. P. 135–140. doi: 10.7326/0003-4819-117-2-135
91. Moons K.G., van Es G.A., Deckers J.W., et al. Limitations of sensitivity, specificity, likelihood ratio, and bayes' theorem in assessing diagnostic probabilities: a clinical example // Epidemiology. 1997. Vol. 8, N 1. P. 12–17. doi: 10.1097/00001648-199701000-00002
92. O'Connor P.W., Tansay C.M., Detsky A.S., et al. The effect of spectrum bias on the utility of magnetic resonance imaging and evoked potentials in the diagnosis of suspected multiple sclerosis // Neurology. 1996. Vol. 47, N 1. P. 140–144. doi: 10.1212/wnl.47.1.140
93. Deckers J.W., Rensing B.J., Tijssen J.G., et al. A comparison of methods of analysing exercise tests for diagnosis of coronary artery disease // Br Heart J. 1989. Vol. 62, N 6. P. 438–444. doi: 10.1136/hrt.62.6.438
94. Naraghi A.M., Gupta S., Jacks L.M., et al. Anterior cruciate ligament reconstruction: MR imaging signs of anterior knee laxity in the presence of an intact graft // Radiology. 2012. Vol. 263, N 3. P. 802–810. doi: 10.1148/radiol.12110779
95. Ashdown H.F., D'Souza N., Karim D., et al. Pain over speed bumps in diagnosis of acute appendicitis: diagnostic accuracy study // BMJ. 2012. Vol. 345, N . P. e8012. doi: 10.1136/bmj.e8012
96. Leeflang M.M., Rutjes A.W., Reitsma J.B., et al. Variation of a test's sensitivity and specificity with disease prevalence // CMAJ. 2013. Vol. 185, N 11. P. E537–544. doi: 10.1503/cmaj.121286
97. Rajaram S., Swift A.J., Capener D., et al. Lung morphology assessment with balanced steady-state free precession MR imaging compared with CT // Radiology. 2012. Vol. 263, N 2. P. 569–577. doi: 10.1148/radiol.12110990
98. Lang T.A., Secic M. Generalizing from a sample to a population: reporting estimates and confidence intervals. Philadelphia: American College of Physicians; 1997.
99. Ioannidis J.P., Evans S.J., Gotzsche P.C., et al. Better reporting of harms in randomized trials: an extension of the CONSORT statement // Ann Intern Med. 2004. Vol. 141, N 10. P. 781–788. doi: 10.7326/0003-4819-141-10-200411160-00009
100. Ioannidis J.P., Lau J. Completeness of safety reporting in randomized trials: an evaluation of 7 medical areas // JAMA. 2001. Vol. 285, N 4. P. 437–443. doi: 10.1001/jama.285.4.437
101. Park S.H., Lee J.H., Lee S.S., et al. CT colonography for detection and characterisation of synchronous proximal colonic lesions in patients with stenosing colorectal cancer // Gut. 2012. Vol. 61, N 12. P. 1716–1722. doi: 10.1136/gutjnl-2011-301135
102. Irwig L.M., Bossuyt P.M., Glasziou P.P., et al. Designing studies to ensure that estimates of test accuracy will travel. In: Knottnerus JA, ed. The evidence base of clinical diagnosis. London: BMJ Publishing Group, 2002. P. 95–116. doi: 10.1002/9781444300574.ch6
103. Ter Riet G., Chesley P., Gross A.G., et al. All that glitters isn't gold: a survey on acknowledgment of limitations in biomedical studies // PLoS ONE. 2013. Vol. 8, N 11. P. e73623. doi: 10.1371/journal.pone.0073623
104. Ioannidis J.P. Limitations are not properly acknowledged in the scientific literature // J Clin Epidemiol. 2007. Vol. 60, N 4. P. 324–329. doi: 10.1016/j.jclinepi.2006.09.011
105. Lord S.J., Irwig L., Simes R.J. When is measuring sensitivity and specificity sufficient to evaluate a diagnostic test, and when do we need randomized trials? // Ann Intern Med. 2006. Vol. 144, N 11. P. 850–855. doi: 10.7326/0003-4819-144-11-200606060-00011
106. Pewsner D., Battaglia M., Minder C., et al. Ruling a diagnosis in or out with 'SpPIn' and 'SnNOut': a note of caution // BMJ. 2004. Vol. 329, N 7459. P. 209–213. doi: 10.1136/bmj.329.7459.209
107. Foerch C., Niessner M., Back T., et al. Diagnostic accuracy of plasma glial fibrillary acidic protein for differentiating intracerebral

hemorrhage and cerebral ischemia in patients with symptoms of acute stroke // Clin Chem. 2012. Vol. 58, N 1. P. 237–245. doi: 10.1373/clinchem.2011.172676
108. Altman D.G. The time has come to register diagnostic and prognostic research // Clin Chem. 2014. Vol. 60, N 4. P. 580–582. doi: 10.1373/clinchem.2013.220335
109. Hooft L., Bossuyt P.M. Prospective registration of marker evaluation studies: time to act // Clin Chem. 2011. Vol. 57, N 12. P. 1684–1686. doi: 10.1373/clinchem.2011.176230
110. Rifai N., Altman D.G., Bossuyt P.M. Reporting bias in diagnostic and prognostic studies: time for action // Clin Chem. 2008. Vol. 54, N 7. P. 1101–1103. doi: 10.1373/clinchem.2008.108993
111. Korevaar D.A., Ochodo E.A., Bossuyt P.M., et al. Publication and reporting of test accuracy studies registered in ClinicalTrials.gov // Clin Chem. 2014. Vol. 60, N 4. P. 651–659. doi: 10.1373/clinchem.2013.218149
112. Rifai N., Bossuyt P.M., Ioannidis J.P., et al. Registering diagnostic and prognostic trials of tests: is it the right thing to do? // Clin Chem. 2014. Vol. 60, N 9. P. 1146–1152. doi: 10.1373/clinchem.2014.226100
113. Korevaar D.A., Bossuyt P.M., Hooft L. Infrequent and incomplete registration of test accuracy studies: analysis of recent study reports // BMJ Open. 2014. Vol. 4, N 1. P. e004596. doi: 10.1136/bmjopen-2013-004596
114. Leeuwenburgh M.M., Wiarda B.M., Wiezer M.J., et al. Comparison of imaging strategies with conditional contrast-enhanced CT and unenhanced MR imaging in patients suspected of having appendicitis: a multicenter diagnostic performance study // Radiology. 2013. Vol. 268, N 1. P. 135–143. doi: 10.1148/radiol.13121753
115. Chan A.W., Song F., Vickers A., et al. Increasing value and reducing waste: addressing inaccessible research // Lancet. 2014. Vol. 383, N 9913. P. 257–266. doi: 10.1016/S0140-6736(13)62296-5
116. Stewart C.M., Schoeman S.A., Booth R.A., et al. Assessment of self taken swabs versus clinician taken swab cultures for diagnosing gonorrhoea in women: single centre, diagnostic accuracy study // BMJ. 2012. Vol. 345. P. e8107. doi: 10.1136/bmj.e8107
117. Sismondo S. Pharmaceutical company funding and its consequences: a qualitative systematic review // Contemp Clin Trials. 2008. Vol. 29, N 2. P. 109–113. doi: 10.1016/j.cct.2007.08.001

REFERENCES

1. Whiting P, Rutjes AW, Reitsma JB, et al. Sources of variation and bias in studies of diagnostic accuracy: a systematic review. *Ann Intern Med*. 2004;140(3):189–202. doi: 10.7326/0003-4819-140-3-200402030-00010
2. Whiting PF, Rutjes AW, Westwood ME, et al. A systematic review classifies sources of bias and variation in diagnostic test accuracy studies. *J Clin Epidemiol*. 2013;66(10):1093–1104. doi: 10.1016/j.jclinepi.2013.05.014
3. Whiting PF, Rutjes AW, Westwood ME, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. *Ann Intern Med*. 2011;155(8):529–536. doi: 10.7326/0003-4819-155-8-201110180-00009
4. Korevaar DA, van Enst WA, Spijker R, et al. Reporting quality of diagnostic accuracy studies: a systematic review and meta-analysis of investigations on adherence to STARD. *Evid Based Med*. 2014;19(2):47–54. doi: 10.1136/eb-2013-101637
5. Korevaar DA, Wang J, van Enst WA, et al. Reporting diagnostic accuracy studies: some improvements after 10 years of STARD. *Radiology*. 2015;274(3):781–789. doi: 10.1148/radiol.14141160
6. Lijmer JG, Mol BW, Heisterkamp S, et al. Empirical evidence of design-related bias in studies of diagnostic tests. *JAMA*. 1999;282(11):1061–1066. doi: 10.1001/jama.282.11.1061
7. Bossuyt PM, Reitsma JB, Bruns DE, et al. Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. *Clin Chem*. 2003;49(7):1–6. doi: 10.1373/49.1.1
8. Begg C, Cho M, Eastwood S, et al. Improving the quality of reporting of randomized controlled trials. The CONSORT statement. *JAMA*. 1996;276(8):637–639. doi: 10.1001/jama.276.8.637
9. Schulz KF, Altman DG, Moher D. CONSORT 2010 statement: updated guidelines for reporting parallel group randomised trials. *BMJ*. 2010;340(1):c332. doi: 10.1136/bmj.c332
10. Bossuyt PM, Reitsma JB, Bruns DE, et al. STARD 2015: an updated list of essential items for reporting diagnostic accuracy studies. *BMJ*. 2015;351:h5527. doi: 10.1136/bmj.h5527
11. Bossuyt PM, Reitsma JB, Bruns DE, et al. The STARD statement for reporting studies of diagnostic accuracy: explanation and elaboration. *Ann Intern Med*. 2003;138(1):W1–12. doi: 10.7326/0003-4819-138-1-200301070-00012-w1
12. Regge D, Laudi C, Galatola G, et al. Diagnostic accuracy of computed tomographic colonography for the detection of advanced neoplasia in individuals at increased risk of colorectal cancer. *JAMA*. 2009;301(23):2453–2461. doi: 10.1001/jama.2009.832
13. Deville WL, Bezemer PD, Bouter LM. Publications on diagnostic test evaluation in family medicine journals: an optimal search strategy. *J Clin Epidemiol*. 2000;53(1):65–69. doi: 10.1016/s0895-4356(99)00144-4
14. Korevaar DA, Cohen JF, Hooft L, et al. Literature survey of high-impact journals revealed reporting weaknesses in abstracts of diagnostic accuracy studies. *J Clin Epidemiol*. 2015;68(6):708–715. doi: 10.1016/j.jclinepi.2015.01.014
15. Korevaar DA, Cohen JF, de Ronde MW, et al. Reporting weaknessess in conference abstracts of diagnostic accuracy studies in ophthalmology. *JAMA Ophthalmol*. 2015;133(12):1464–1467. doi: 10.1001/jamaophthalmol.2015.3577
16. A proposal for more informative abstracts of clinical articles. Ad Hoc Working Group for Critical Appraisal of the Medical Literature. *Ann Intern Med*. 1987;106(4):598–604.
17. Stiell IG, Greenberg GH, Wells GA, et al. Derivation of a decision rule for the use of radiography in acute knee injuries. *Ann Emerg Med*. 1995;26(4):405–413. doi: 10.1016/s0196-0644(95)70106-0
18. Horvath AR, Lord SJ, StJohn A, et al. From biomarkers to medical tests: the changing landscape of test evaluation. *Clin Chim Acta*. 2014;427:49–57. doi: 10.1016/j.cca.2013.09.018
19. Bossuyt PM, Irwig L, Craig J, et al. Comparative accuracy: assessing new tests against existing diagnostic pathways. *BMJ*. 2006;332:1089–1092. doi: 10.1136/bmj.332.7549.1089
20. Gieseker KE, Roe MH, MacKenzie T, et al. Evaluating the American Academy of Pediatrics diagnostic standard for Streptococcus pyogenes pharyngitis: backup culture versus repeat rapid antigen testing. *Pediatrics*. 2003;111(6 Pt 1):e666–670. doi: 10.1542/peds.111.6.e666
21. Tanz RR, Gerber MA, Kabat W, et al. Performance of a rapid antigen-detection test and throat culture in community pediatric

offices: implications for management of pharyngitis. *Pediatrics*. 2009;123(2):437–444. doi: 10.1542/peds.2008-0488
22. Ochodo EA, de Haan MC, Reitsma JB, et al. Overinterpretation and misreporting of diagnostic accuracy studies: evidence of 'spin'. *Radiology*. 2013;267(2):581–588. doi: 10.1148/radiol.12120527
23. Freer PE, Niell B, Rafferty EA. Preoperative tomosynthesis-guided needle localization of mammographically and sonographically occult breast lesions. *Radiology*. 2015;275(2):377–383. doi: 10.1148/radiol.14140515
24. Sorensen HT, Sabroe S, Olsen J. A framework for evaluation of secondary data sources for epidemiological research. *Int J Epidemiol*. 1996;25(2):435–442. doi: 10.1093/ije/25.2.435
25. Geersing GJ, Erkens PM, Lucassen WA, et al. Safe exclusion of pulmonary embolism using the Wells rule and qualitative D-dimer testing in primary care: prospective cohort study. *BMJ*. 2012;345:e6564. doi: 10.1136/bmj.e6564
26. Bomers MK, van Agtmael MA, Luik H, et al. Using a dog's superior olfactory sensitivity to identify Clostridium difficile in stools and patients: proof of principle study. *BMJ*. 2012;345:e7396. doi: 10.1136/bmj.e7396
27. Philbrick JT, Horwitz RI, Feinstein AR. Methodologic problems of exercise testing for coronary artery disease: groups, analysis and bias. *Am J Cardiol*. 1980;46(5):807–812. doi: 10.1016/0002-9149(80)90432-4
28. Rutjes AW, Reitsma JB, Vandenbroucke JP, et al. Case-control and two-gate designs in diagnostic accuracy studies. *Clin Chem*. 2005;51(8):1335–1341. doi: 10.1373/clinchem.2005.048595
29. Rutjes AW, Reitsma JB, Di Nisio M, et al. Evidence of bias and variation in diagnostic accuracy studies. *CMAJ*. 2006;174(4):469–476. doi: 10.1503/cmaj.050090
30. Knottnerus JA, Muris JW. Assessment of the accuracy of diagnostic tests: the cross-sectional study. *J Clin Epidemiol*. 2003;56(11):1118–1128. doi: 10.1016/s0895-4356(03)00206-3
31. van der Schouw YT, Van Dijk R, Verbeek AL. Problems in selecting the adequate patient population from existing data files for assessment studies of new diagnostic tests. *J Clin Epidemiol*. 1995;48(3):417–422. doi: 10.1016/0895-4356(94)00144-f
32. Leeflang MM, Bossuyt PM, Irwig L. Diagnostic test accuracy may vary with prevalence: implications for evidence-based diagnosis. *J Clin Epidemiol*. 2009;62(1):5–12. doi: 10.1016/j.jclinepi.2008.04.007
33. Attia M, Zaoutis T, Eppes S, et al. Multivariate predictive models for group A beta-hemolytic streptococcal pharyngitis in children. *Acad Emerg Med*. 1999;6(1):8–13. doi: 10.1111/j.1553-2712.1999.tb00087.x
34. Knottnerus JA, Knipschild PG, Sturmans F. Symptoms and selection bias: the influence of selection towards specialist care on the relationship between symptoms and diagnoses. *Theor Med*. 1989;10(1):67–81. doi: 10.1007/BF00625761
35. Knottnerus JA, Leffers P. The influence of referral patterns on the characteristics of diagnostic tests. *J Clin Epidemiol*. 1992;45(10):1143–1154. doi: 10.1016/0895-4356(92)90155-g
36. Melbye H, Straume B. The spectrum of patients strongly influences the usefulness of diagnostic tests for pneumonia. *Scand J Prim Health Care*. 1993;11:241–246. doi: 10.3109/02813439308994838
37. Ezike EN, Rongkavilit C, Fairfax MR, et al. Effect of using 2 throat swabs vs 1 throat swab on detection of group A streptococcus by a rapid antigen detection test. *Arch Pediatr Adolesc Med*. 2005;159(5):486–490. doi: 10.1001/archpedi.159.5.486
38. Rosjo H, Kravdal G, Hoiseth AD, et al. Troponin I measured by a high-sensitivity assay in patients with suspected reversible myocardial ischemia: data from the Akershus Cardiac Examination (ACE) 1 study. *Clin Chem*. 2012;58(11):1565–1573. doi: 10.1373/clinchem.2012.190868
39. Irwig L, Bossuyt P, Glasziou P, et al. Designing studies to ensure that estimates of test accuracy are transferable. *BMJ*. 2002;324(7338):669–671. doi: 10.1136/bmj.324.7338.669
40. Detrano R, Gianrossi R, Froelicher V. The diagnostic accuracy of the exercise electrocardiogram: a meta-analysis of 22 years of research. *Prog Cardiovasc Dis*. 1989;32(3):173–206. doi: 10.1016/0033-0620(89)90025-x
41. Brealey S, Scally AJ. Bias in plain film reading performance studies. *Br J Radiol*. 2001;74(880):307–316. doi: 10.1259/bjr.74.880.740307
42. Elmore JG, Wells CK, Lee CH, et al. Variability in radiologists' interpretations of mammograms. *N Engl J Med*. 1994;331(22):1493–1499. doi: 10.1056/NEJM199412013312206
43. Ronco G, Montanari G, Aimone V, et al. Estimating the sensitivity of cervical cytology: errors of interpretation and test limitations. *Cytopathology*. 1996;7(3):151–158. doi: 10.1046/j.1365-2303.1996.39382393.x
44. Cohen MB, Rodgers RP, Hales MS, et al. Influence of training and experience in fine-needle aspiration biopsy of breast. Receiver operating characteristics curve analysis. *Arch Pathol Lab Med*. 1987;111(6):518–520.
45. Fox JW, Cohen DM, Marcon MJ, et al. Performance of rapid streptococcal antigen testing varies by personnel. *J Clin Microbiol*. 2006;44(11):3918–3922. doi: 10.1128/JCM.01399-06
46. Gandy M, Sharpe L, Perry KN, et al. Assessing the efficacy of 2 screening measures for depression in people with epilepsy. *Neurology*. 2012;79(4):371–375. doi: 10.1212/WNL.0b013e318260cbfc
47. Stegeman I, de Wijkerslooth TR, Stoop EM, et al. Combining risk factors with faecal immunochemical test outcome for selecting CRC screenees for colonoscopy. *Gut*. 2014;63(3):466–471. doi: 10.1136/gutjnl-2013-305013
48. Leeflang MM, Moons KG, Reitsma JB, et al. Bias in sensitivity and specificity caused by data-driven selection of optimal cutoff values: mechanisms, magnitude, and solutions. *Clin Chem*. 2008;54(4):729–737. doi: 10.1373/clinchem.2007.096032
49. Ewald B. Post hoc choice of cut points introduced bias to diagnostic research. *J Clin Epidemiol*. 2006;59(8):798–801. doi: 10.1016/j.jclinepi.2005.11.025
50. Justice AC, Covinsky KE, Berlin JA. Assessing the generalizability of prognostic information. *Ann Intern Med*. 1999;130(6):515–524. doi: 10.7326/0003-4819-130-6-199903160-00016
51. Harrell FE Jr., Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat Med*. 1996;15(4):361–387. doi: 10.1002/(SICI)1097-0258(19960229)15:4<361::AID-SIM168>3.0.CO;2-4
52. Hodgdon T, McInnes MD, Schieda N, et al. Can quantitative CT texture analysis be used to differentiate fat-poor renal angiomyolipoma from renal cell carcinoma on unenhanced CT images? *Radiology*. 2015;276(3):787–796. doi: 10.1148/radiol.2015142215
53. Begg CB. Biases in the assessment of diagnostic tests. *Stat Med*. 1987;6(4):411–423. doi: 10.1002/sim.4780060402
54. Doubilet P, Herman PG. Interpretation of radiographs: effect of clinical history. *AJR Am J Roentgenol*. 1981;137(5):1055–1058. doi: 10.2214/ajr.137.5.1055

55. D'Orsi CJ, Getty DJ, Pickett RM, et al. Stereoscopic digital mammography: improved specificity and reduced rate of recall in a prospective clinical trial. *Radiology*. 2013;266(1):81–88. doi: 10.1148/radiol.12120382
56. Knottnerus JA, Buntinx F. The evidence base of clinical diagnosis: theory and methods of diagnostic research. 2nd edn. BMJ Books, 2008. 316 p.
57. Pepe M. Study design and hypothesis testing. The statistical evaluation of medical tests for classification and prediction. Oxford, UK: Oxford University Press; 2003. P. 214–251.
58. Hayen A, Macaskill P, Irwig L, et al. Appropriate statistical methods are required to assess diagnostic tests for replacement, add-on, and triage. *J Clin Epidemiol*. 2010;63(8):883–891. doi: 10.1016/j.jclinepi.2009.08.024
59. Garcia Pena BM, Mandl KD, Kraus SJ, et al. Ultrasonography and limited computed tomography in the diagnosis and management of appendicitis in children. *JAMA*. 1999;282(11):1041–1046. doi: 10.1001/jama.282.11.1041
60. Simel DL, Feussner JR, DeLong ER, et al. Intermediate, indeterminate, and uninterpretable diagnostic test results. *Med Decis Making*. 1987;7(2):107–114. doi: 10.1177/0272989X8700700208
61. Philbrick JT, Horwitz RI, Feinstein AR, et al. The limited spectrum of patients studied in exercise test research. Analyzing the tip of the iceberg. *JAMA*. 1982;248(19):2467–2470.
62. Begg CB, Greenes RA, Iglewicz B. The influence of uninterpretability on the assessment of diagnostic tests. *J Chronic Dis*. 1986;39(8):575–584. doi: 10.1016/0021-9681(86)90182-7
63. Shinkins B, Thompson M, Mallett S, et al. Diagnostic accuracy studies: how to report and analyse inconclusive test results. *BMJ*. 2013;346:f2778. doi: 10.1136/bmj.f2778
64. Pisano ED, Fajardo LL, Tsimikas J, et al. Rate of insufficient samples for fine-needle aspiration for nonpalpable breast lesions in a multicenter clinical trial: the Radiologic Diagnostic Oncology Group 5 Study. The RDOG5 investigators. *Cancer*. 1998;82(4):679–688. doi: 10.1002/(sici)1097-0142(19980215)82:4<679::aid-cncr10>3.0.co;2-v
65. Giard RW, Hermans J. The value of aspiration cytologic examination of the breast. A statistical review of the medical literature. *Cancer*. 1992;69(8):2104–2110. doi: 10.1002/1097-0142(19920415)69:8<2104::aid-cncr2820690816>3.0.co;2-o
66. Investigators P. Value of the ventilation/perfusion scan in acute pulmonary embolism. Results of the prospective investigation of pulmonary embolism diagnosis (PIOPED). *JAMA*. 1990;263(20):2753–2759. doi: 10.1001/jama.1990.03440200057023
67. Min JK, Leipsic J, Pencina MJ, et al. Diagnostic accuracy of fractional flow reserve from anatomic CT angiography. *JAMA*. 2012;308(12):1237–1245. doi: 10.1001/2012.jama.11274
68. Naaktgeboren CA, de Groot JA, Rutjes AW, et al. Anticipating missing reference standard data when planning diagnostic accuracy studies. *BMJ*. 2016;352:i402. doi: 10.1136/bmj.i402
69. Van der Heijden GJ, Donders AR, Stijnen T, et al. Imputation of missing values is superior to complete case analysis and the missing-indicator method in multivariable diagnostic research: a clinical example. *J Clin Epidemiol*. 2006;59(10):1102–1109. doi: 10.1016/j.jclinepi.2006.01.015
70. de Groot JA, Bossuyt PM, Reitsma JB, et al. Verification problems in diagnostic accuracy studies: consequences and solutions. *BMJ*. 2011;343:d4770. doi: 10.1136/bmj.d4770
71. Pons B, Lautrette A, Oziel J, et al. Diagnostic accuracy of early urinary index changes in differentiating transient from persistent acute kidney injury in critically ill patients: multicenter cohort study. *Crit Care*. 2013;17(2):R56. doi: 10.1186/cc12582
72. Sun X, Ioannidis JP, Agoritsas T, et al. How to use a subgroup analysis: users' guide to the medical literature. *JAMA*. 2014;311(4):405–411. doi: 10.1001/jama.2013.285063
73. Zalis ME, Blake MA, Cai W, et al. Diagnostic accuracy of laxative-free computed tomographic colonography for detection of adenomatous polyps in asymptomatic adults: a prospective evaluation. *Ann Intern Med*. 2012;156(10):692–702. doi: 10.7326/0003-4819-156-10-201205150-00005
74. Flahault A, Cadilhac M, Thomas G. Sample size calculation should be performed for design accuracy in diagnostic test studies. *J Clin Epidemiol*. 2005;58(8):859–862. doi: 10.1016/j.jclinepi.2004.12.009
75. Pepe MS. The statistical evaluation of medical tests for classification and prediction. Oxford, New York: Oxford University Press, 2003.
76. Vach W, Gerke O, Hoilund-Carlsen PF. Three principles to define the success of a diagnostic study could be identified. *J Clin Epidemiol*. 2012;65(3):293–300. doi: 10.1016/j.jclinepi.2011.07.004
77. Bachmann LM, Puhan MA, ter Riet G, et al. Sample sizes of studies on diagnostic accuracy: literature survey. *BMJ*. 2006;332(4550):1127–1129. doi: 10.1136/bmj.38793.637789.2F
78. Bochmann F, Johnson Z, Azuara-Blanco A. Sample size in studies on diagnostic accuracy in ophthalmology: a literature survey. *Br J Ophthalmol*. 2007;91(7):898–900. doi: 10.1136/bjo.2006.113290
79. Collins MG, Teo E, Cole SR, et al. Screening for colorectal cancer and advanced colorectal neoplasia in kidney transplant recipients: cross sectional prevalence and diagnostic accuracy study of faecal immunochemical testing for haemoglobin and colonoscopy. *BMJ*. 2012;345:e4657. doi: 10.1136/bmj.e4657
80. Cecil MP, Kosinski AS, Jones MT, et al. The importance of work-up (verification) bias correction in assessing the accuracy of SPECT thallium-201 testing for the diagnosis of coronary artery disease. *J Clin Epidemiol*. 1996;49(7):735–742. doi: 10.1016/0895-4356(96)00014-5
81. Choi BC. Sensitivity and specificity of a single diagnostic test in the presence of work-up bias. *J Clin Epidemiol*. 1992;45(6):581–586. doi: 10.1016/0895-4356(92)90129-b
82. Diamond GA. Off Bayes: effect of verification bias on posterior probabilities calculated using Bayes' theorem. *Med Decis Making*. 1992;12(1):22–31. doi: 10.1177/0272989X9201200105
83. Diamond GA, Rozanski A, Forrester JS, et al. A model for assessing the sensitivity and specificity of tests subject to selection bias. Application to exercise radionuclide ventriculography for diagnosis of coronary artery disease. *J Chronic Dis*. 1986;39(5):343–355. doi: 10.1016/0021-9681(86)90119-0
84. Greenes RA, Begg CB. Assessment of diagnostic technologies. Methodology for unbiased estimation from samples of selectively verified patients. *Invest Radiol*. 1985;20(7):751–756.
85. Ransohoff DF, Feinstein AR. Problems of spectrum and bias in evaluating the efficacy of diagnostic tests. *N Engl J Med*. 1978;299(17):926–930. doi: 10.1056/NEJM197810262991705
86. Zhou XH. Effect of verification bias on positive and negative predictive values. *Stat Med*. 1994;13(17):1737–1745. doi: 10.1002/sim.4780131705
87. Kok L, Elias SG, Witteman BJ, et al. Diagnostic accuracy of point-of-care fecal calprotectin and immunochemical occult blood

tests for diagnosis of organic bowel disease in primary care: the Cost-Effectiveness of a Decision Rule for Abdominal Complaints in Primary Care (CEDAR) study. *Clin Chem*. 2012;58(6):989–998. doi: 10.1373/clinchem.2011.177980

88. Harris JM. The hazards of bedside Bayes. *JAMA*. 1981;246(22):2602–2605.

89. Hlatky MA, Pryor DB, Harrell FE, et al. Factors affecting sensitivity and specificity of exercise electrocardiography. Multivariable analysis. *Am J Med*. 1984;77(1):64–71. doi: 10.1016/0002-9343(84)90437-6

90. Lachs MS, Nachamkin I, Edelstein PH, et al. Spectrum bias in the evaluation of diagnostic tests: lessons from the rapid dipstick test for urinary tract infection. *Ann Intern Med*. 1992;117(2):135–140. doi: 10.7326/0003-4819-117-2-135

91. Moons KG, van Es GA, Deckers JW, et al. Limitations of sensitivity, specificity, likelihood ratio, and bayes' theorem in assessing diagnostic probabilities: a clinical example. *Epidemiology*. 1997;8(1):12–17. doi: 10.1097/00001648-199701000-00002

92. O'Connor PW, Tansay CM, Detsky AS, et al. The effect of spectrum bias on the utility of magnetic resonance imaging and evoked potentials in the diagnosis of suspected multiple sclerosis. *Neurology*. 1996;47(1):140–144. doi: 10.1212/wnl.47.1.140

93. Deckers JW, Rensing BJ, Tijssen JG, et al. A comparison of methods of analysing exercise tests for diagnosis of coronary artery disease. *Br Heart J*. 1989;62(6):438–444. doi: 10.1136/hrt.62.6.438

94. Naraghi AM, Gupta S, Jacks LM, et al. Anterior cruciate ligament reconstruction: MR imaging signs of anterior knee laxity in the presence of an intact graft. *Radiology*. 2012;263(3):802–810. doi: 10.1148/radiol.12110779

95. Ashdown HF, D'Souza N, Karim D, et al. Pain over speed bumps in diagnosis of acute appendicitis: diagnostic accuracy study. *BMJ*. 2012;345:e8012. doi: 10.1136/bmj.e8012

96. Leeflang MM, Rutjes AW, Reitsma JB, et al. Variation of a test's sensitivity and specificity with disease prevalence. *CMAJ*. 2013;185(11):E537–544. doi: 10.1503/cmaj.121286

97. Rajaram S, Swift AJ, Capener D, et al. Lung morphology assessment with balanced steady-state free precession MR imaging compared with CT. *Radiology*. 2012;263(2):569–577. doi: 10.1148/radiol.12110990

98. Lang TA, Secic M. Generalizing from a sample to a population: reporting estimates and confidence intervals. Philadelphia: American College of Physicians; 1997.

99. Ioannidis JP, Evans SJ, Gotzsche PC, et al. Better reporting of harms in randomized trials: an extension of the CONSORT statement. *Ann Intern Med*. 2004;141(10):781–788. doi: 10.7326/0003-4819-141-10-200411160-00009

100. Ioannidis JP, Lau J. Completeness of safety reporting in randomized trials: an evaluation of 7 medical areas. *JAMA*. 2001;285(4):437–443. doi: 10.1001/jama.285.4.437

101. Park SH, Lee JH, Lee SS, et al. CT colonography for detection and characterisation of synchronous proximal colonic lesions in patients with stenosing colorectal cancer. *Gut*. 2012;61(12):1716–1722. doi: 10.1136/gutjnl-2011-301135

102. Irwig LM, Bossuyt PM, Glasziou PP, et al. Designing studies to ensure that estimates of test accuracy will travel. In: Knottnerus JA, ed. The evidence base of clinical diagnosis. London: BMJ Publishing Group; 2002. P. 95–116. doi: 10.1002/9781444300574.ch6

103. Ter Riet G, Chesley P, Gross AG, et al. All that glitters isn't gold: a survey on acknowledgment of limitations in biomedical studies. *PLoS ONE*. 2013;8(11):e73623. doi: 10.1371/journal.pone.0073623

104. Ioannidis JP. Limitations are not properly acknowledged in the scientific literature. *J Clin Epidemiol*. 2007;60(4):324–329. doi: 10.1016/j.jclinepi.2006.09.011

105. Lord SJ, Irwig L, Simes RJ. When is measuring sensitivity and specificity sufficient to evaluate a diagnostic test, and when do we need randomized trials? *Ann Intern Med*. 2006;144(11):850–855. doi: 10.7326/0003-4819-144-11-200606060-00011

106. Pewsner D, Battaglia M, Minder C, et al. Ruling a diagnosis in or out with 'SpPIn' and 'SnNOut': a note of caution. *BMJ*. 2004;329(7459):209–213. doi: 10.1136/bmj.329.7459.209

107. Foerch C, Niessner M, Back T, et al. Diagnostic accuracy of plasma glial fibrillary acidic protein for differentiating intracerebral hemorrhage and cerebral ischemia in patients with symptoms of acute stroke. *Clin Chem*. 2012;58(1):237–245. doi: 10.1373/clinchem.2011.172676

108. Altman DG. The time has come to register diagnostic and prognostic research. *Clin Chem*. 2014;60(4):580–582. doi: 10.1373/clinchem.2013.220335

109. Hooft L, Bossuyt PM. Prospective registration of marker evaluation studies: time to act. *Clin Chem*. 2011;57(12):1684–1686. doi: 10.1373/clinchem.2011.176230

110. Rifai N, Altman DG, Bossuyt PM. Reporting bias in diagnostic and prognostic studies: time for action. *Clin Chem*. 2008;54(7):1101–1103. doi: 10.1373/clinchem.2008.108993

111. Korevaar DA, Ochodo EA, Bossuyt PM, et al. Publication and reporting of test accuracy studies registered in ClinicalTrials.gov. *Clin Chem*. 2014;60(4):651–659. doi: 10.1373/clinchem.2013.218149

112. Rifai N, Bossuyt PM, Ioannidis JP, et al. Registering diagnostic and prognostic trials of tests: is it the right thing to do? *Clin Chem*. 2014;60(9):1146–1152. doi: 10.1373/clinchem.2014.226100

113. Korevaar DA, Bossuyt PM, Hooft L. Infrequent and incomplete registration of test accuracy studies: analysis of recent study reports. *BMJ Open*. 2014;4(1):e004596. doi: 10.1136/bmjopen-2013-004596

114. Leeuwenburgh MM, Wiarda BM, Wiezer MJ, et al. Comparison of imaging strategies with conditional contrast-enhanced CT and unenhanced MR imaging in patients suspected of having appendicitis: a multicenter diagnostic performance study. *Radiology*. 2013;268(1):135–143. doi: 10.1148/radiol.13121753

115. Chan AW, Song F, Vickers A, et al. Increasing value and reducing waste: addressing inaccessible research. *Lancet*. 2014;383(9913):257–266. doi: 10.1016/S0140-6736(13)62296-5

116. Stewart CM, Schoeman SA, Booth RA, et al. Assessment of self taken swabs versus clinician taken swab cultures for diagnosing gonorrhoea in women: single centre, diagnostic accuracy study. *BMJ*. 2012;345:e8107. doi: 10.1136/bmj.e8107

117. Sismondo S. Pharmaceutical company funding and its consequences: a qualitative systematic review. *Contemp Clin Trials*. 2008;29(2):109–113. doi: 10.1016/j.cct.2007.08.001

ОБ АВТОРАХ

*** Patrick M.M. Bossuyt,** Department of Clinical Epidemiology, Biostatistics and Bioinformatics, Academic Medical Centre, University of Amsterdam, Амстердам, Нидерланды;
ORCID: https://orcid.org/0000-0003-4427-0128;
e-mail: p.m.bossuyt@amc.uva.nl

Jérémie F. Cohen; Epidemiology, Biostatistics and Bioinformatics, Academic Medical Centre, University of Amsterdam, Амстердам, Нидерланды; Department of Pediatrics, INSERM UMR 1153, Necker Hospital, AP-HP, Paris Descartes University, Париж, Франция;
ORCID: https://orcid.org/0000-0003-3572-8985

Daniël A. Korevaar; Department of Clinical Epidemiology, Biostatistics and Bioinformatics, Academic Medical Centre, University of Amsterdam, Амстердам, Нидерланды;
ORCID: https://orcid.org/0000-0002-7979-7897

Douglas G. Altman; Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences, Centre for Statistics in Medicine, University of Oxford, Оксфорд, Великобритания;
ORCID: https://orcid.org/0000-0002-7183-4083

David E. Bruns; Department of Pathology, University of Virginia School of Medicine, Шарлотсвилл, Вирджиния, США

Constantine A. Gatsonis; Department of Biostatistics, Brown University School of Public Health, Провиденс, Род-Айленд, США

Lotty Hooft; Cochrane Netherlands, Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, University of Utrecht, Утрехт, Нидерланды

Les Irwig; Screening and Diagnostic Test Evaluation Program, School of Public Health, University of Sydney, Сидней, Новый Южный Уэльс, Австралия

Deborah Levine; Department of Radiology, Beth Israel Deaconess Medical Center, Бостон, Массачусетс, США; Radiology Editorial Office, Бостон, Массачусетс, США;
ORCID: https://orcid.org/0000-0001-7761-6493

Johannes B. Reitsma; Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, University of Utrecht, Утрехт, Нидерланды

Henrica C.W. de Vet; Department of Epidemiology and Biostatistics, EMGO Institute for Health and Care Research, VU University Medical Center, Амстердам, Нидерланды;
ORCID: https://orcid.org/0000-0002-5454-2804

AUTHORS' INFO

*** Patrick M.M. Bossuyt,** Department of Clinical Epidemiology, Biostatistics and Bioinformatics, Academic Medical Centre, University of Amsterdam, Amsterdam, The Netherlands;
ORCID: https://orcid.org/0000-0003-4427-0128;
e-mail: p.m.bossuyt@amc.uva.nl

Jérémie F. Cohen; Epidemiology, Biostatistics and Bioinformatics, Academic Medical Centre, University of Amsterdam, Amsterdam, The Netherlands; Department of Pediatrics, INSERM UMR 1153, Necker Hospital, AP-HP, Paris Descartes University, Paris, France;
ORCID: https://orcid.org/0000-0003-3572-8985

Daniël A. Korevaar; Department of Clinical Epidemiology, Biostatistics and Bioinformatics, Academic Medical Centre, University of Amsterdam, Amsterdam, The Netherlands;
ORCID: https://orcid.org/0000-0002-7979-7897

Douglas G. Altman; Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences, Centre for Statistics in Medicine, University of Oxford, Oxford, UK;
ORCID: https://orcid.org/0000-0002-7183-4083

David E. Bruns; Department of Pathology, University of Virginia School of Medicine, Charlottesville, Virginia, USA

Constantine A. Gatsonis; Department of Biostatistics, Brown University School of Public Health, Providence, Rhode Island, USA

Lotty Hooft; Cochrane Netherlands, Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, University of Utrecht, Utrecht, The Netherlands

Les Irwig; Screening and Diagnostic Test Evaluation Program, School of Public Health, University of Sydney, Sydney, New South Wales, Australia

Deborah Levine; Department of Radiology, Beth Israel Deaconess Medical Center, Boston, Massachusetts, USA; Radiology Editorial Office, Boston, Massachusetts, USA;
ORCID: https://orcid.org/0000-0001-7761-6493

Johannes B. Reitsma; Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, University of Utrecht, Utrecht, The Netherlands

Henrica C.W. de Vet; Department of Epidemiology and Biostatistics, EMGO Institute for Health and Care Research, VU University Medical Center, Amsterdam, The Netherlands;
ORCID: https://orcid.org/0000-0002-5454-2804

* Автор, ответственный за переписку / Corresponding author