NB: Версия, принятая Всемирной ассамблеей делегатов МЭБ в мае 2014 года

ГЛАВА 2.6.5.

СТАТИСТИЧЕСКИЕ ПОДХОДЫ К ВАЛИДАЦИИ

ВВЕДЕНИЕ

Рекомендации МЭБ по валидации содержат подробную информацию и примеры, подтверждающие содержание Стандарта валидации МЭБ, опубликованного в форме главы 1.1.6 Руководства по диагностическим тестам и вакцинам для наземных животных или главы 1.1.2 Руководства по диагностическим тестам для водных животных. Термин «Стандарт валидации МЭБ», используемый в настоящей главе, следует понимать как отсылку к указанным главам.

Выбор статистических методов анализа данных валидации тестов и испытаний, полученных в процессе лабораторных экспериментов и оценки полевых образцов, зависит от таких факторов, как план исследования и отбор образцов (источник, число образцов, число повторных анализов/тестов и т. д.). Конкретные указания, касающиеся «лучшего подхода», следует давать после консультации со специалистом по статистике и на этапе планирования, перед началом валидационных исследований.

В целях краткости изложения, в настоящем приложении рассматриваются общепринятые подходы к валидации потенциального теста/испытания, а, следовательно, не учитываются все статистические методы, которые могли бы использоваться на практике. Различные методы описываются с целью оценки точности анализа при многократном повторении (повторяемость и воспроизводимость), диагностическую чувствительность (ДЧ) и диагностическую специфичность (ДС) аналитических характеристик (диагностическую чувствительность и диагностическую специфичность), а также диагностические характеристики (например, область ниже кривой операционной характеристики приемника (кривой зависимости чувствительности от частоты ложноположительных заключений)), используемые для обнаружения анализируемого материала у отдельных животных. Аналогичные принципы применяются при использовании испытаний/тестов с целью обнаружения идентичного анализируемого материала в естественных или искусственно созданных пулах образцов, полученных у животных в группах (например, в стадах или отарах). В этом случае единицей эпидемиологического исследования являются не столько отдельные животные, сколько группы животных.

Определения шкал измерения:

Двоичная (дихотомическая) шкала: Либо положительная, либо отрицательная, поскольку это связано с тем, как представлены результаты исследования, или положительная/отрицательная при выбранном пороговом (критическом) значении, когда результаты измеряются по порядковой (одинарной) или непрерывной (числовой) шкале.

Порядковая шкала: Измерения производятся по шкале с дискретными значениями, где более высокие значения обычно обозначают наличие большего числа анализируемого материала (вещества), например титры сыворотки, нейтрализующие вирус.

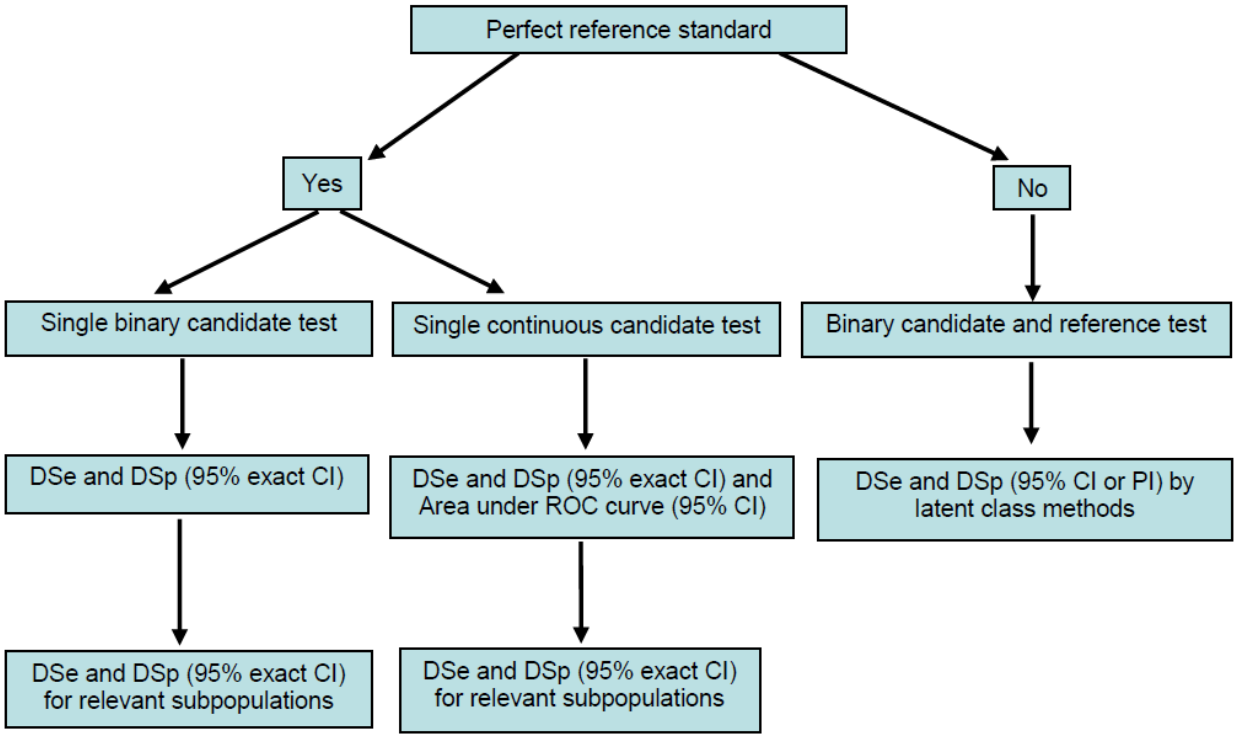
Непрерывная шкала: Бесконечное число измеряемых значений теоретически возможно, в зависимости от системы измерения, например, оптической плотности или процента положительных значений в твердофазном иммуноферментном анализе, а также циклических пороговых значений, полученных с помощью реакции твердофазных иммуноферментных анализов в масштабе реального времени, которые ниже максимального числа циклов анализа.

Статистические методы различаются в зависимости от того, оцениваются ли однократные или многократные испытания, их шкал измерения (двоичной, порядковой или непрерывной), используются ли независимые или зависимые (парные) выборки (образцы), и имеется ли абсолютно точный эталонный стандарт (часто именуемый «золотым стандартом») для

сравнения (Wilks, 2001). Блок-схемы выбора статистических методов оценки измерений диагностической точности, таких как диагностическая чувствительность и диагностическая специфичность, представлены на рисунках 1 и 2.

Соответствие плана исследования и статистического анализа может не всегда отражаться на качестве отчетов, представляемых в научных публикациях, в связи с чем поощряется следование разработчиками и специалистами по оценке испытаний типовому контрольному перечню СОДТ (Стандартов отчета о диагностической точности) (Bossuyt et al., 2003), позволяющему удостовериться в полноте представления соответствующей информации при проведении валидационных исследований инфекционных болезней у животных.

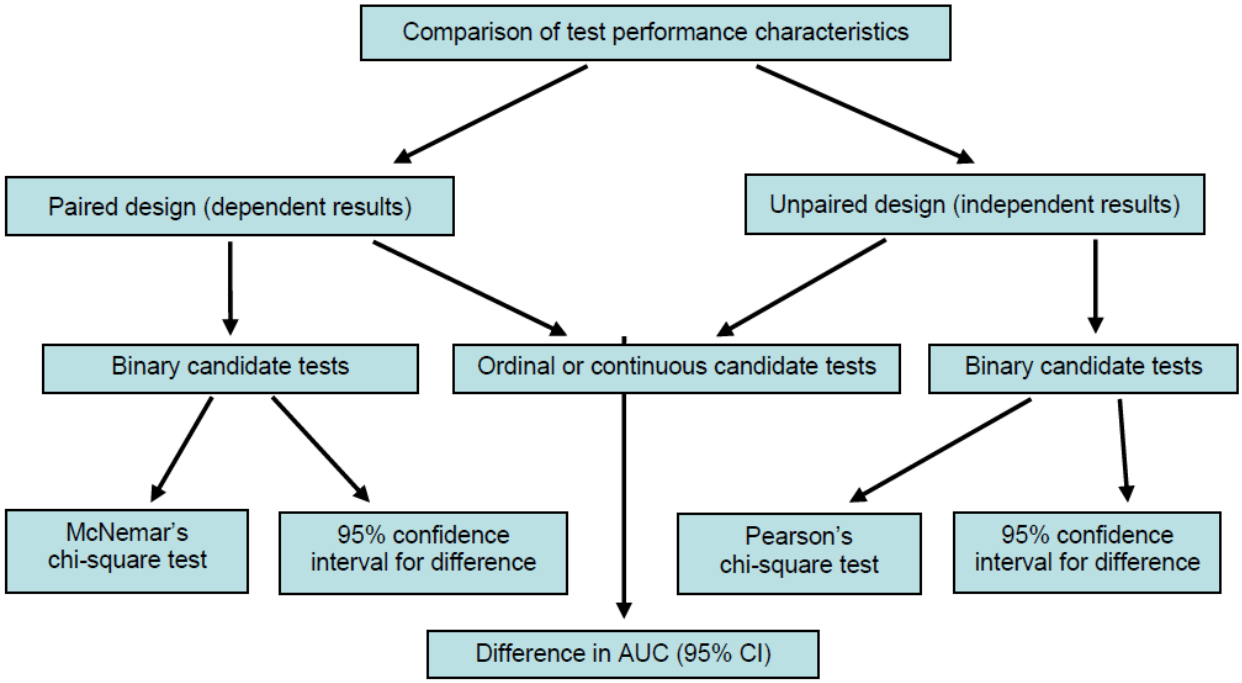
Инструкцию по проведению анализа данных о погрешностях измерений и информацию о сравнительных анализах методов смотрите в главах 3.6.4 и 3.6.8, соответственно.



Perfect reference standard	Эталонный стандарт
Yes	Да
No	Нет
Single binary candidate test	Однократное потенциальное испытание с использованием двоичной шкалы
Single continuous candidate test	Однократное потенциальное испытание с использованием непрерывной шкалы
Binary candidate and reference test	Потенциальное испытание с использованием двоичной шкалы и эталонное испытание
DSe and DSp (95% exact CI)	ДЧ и ДС (95%)
DSe and DSp (95% exact CI) and Area under ROC curve (95% CI)	ДЧ и ДС (95% точного ДИ) и область ниже кривой ОХП (95% ДИ)
DSe and DSp (95% CI or PI) by latent class methods	ДЧ и ДС (95% ДИ или ИВ), определенные методом анализа латентных классов
DSe and DSp (95% exact CI) for relevant subpopulations	ДЧ и ДС (95% точного ДИ) для соответствующих субпопуляций

Аббревиатуры: ДЧ = диагностическая чувствительность; ДС = диагностическая специфичность; ОХП = операционная характеристика приемника (зависимость чувствительности от частоты ложноположительных заключений); ДИ = доверительный интервал; ИВ = интервал вероятности

Рисунок 1. *Блок-схема предлагаемых методов статистического анализа при оценке однократного потенциального испытания с использованием эталонного стандарта и без использования такового*



Comparison of test performance characteristics	Сравнение характеристик испытания
Paired design (dependent results)	План испытания с парными выборками (зависимыми результатами)
Unpaired design (independent results)	План испытания с непарными выборками (независимыми результатами)
Binary candidate tests	Потенциальные испытания с использованием двоичной шкалы
Ordinal or continuous candidate tests	Потенциальные испытания с использованием порядковой или непрерывной шкалы
Binary candidate tests	Потенциальные испытания с использованием двоичной шкалы
McNemar's chi-square test	Критерий хи-квадрат МакНемара
95% confidence interval for difference	95% доверительного интервала для разницы
Pearson's chi-square test	Критерий хи-квадрат Пирсона
Difference in AUC (95% CI)	Разница в ОНК (95% ДИ)

Abbreviation: ОНК = область ниже кривой операционной характеристики приемника

Рисунок 2. *Блок-схема предлагаемых методов статистического анализа при оценке показателей диагностической чувствительности (ДЧ) и диагностической специфичности (ДС), а также ОНК многократных испытаний с использованием эталонного стандарта. Порядковые и непрерывные данные должны анализироваться в их изначальной форме и в качестве результатов по двоичной шкале при рекомендованном пороговом значении. Анализы должны проводиться, как для оценки данных ДЧ, так и для оценки данных ДС, при которой эти данные доступны.*

1. Воспроизводимость испытания в одной лаборатории

Оценка воспроизводимости испытания в одной лаборатории (часто определяемая термином *точность*, когда измерения производятся с использованием непрерывной шкалы) требует наличия как минимум в трех образцах таких концентраций анализируемого материала, находящихся в

пределах рабочего диапазона испытания, которые проверялись бы с помощью повторных испытаний единственным оператором с использованием серии или партии комплектов однократных испытаний (тестов). Как правило, такие испытания проводятся в один день, однако проведение их в разные дни также возможно. Использование трех или четырех образцов для повторных испытаний вместо двух образцов является более предпочтительным, потому что такой способ позволяет лучше передать внутреннюю изменчивость (вариацию) результатов испытаний в процессе их проведения. Учитывая расходы, использование более двух образцов для повторных испытаний может оказаться неоправданным для всех типов испытаний (например, для испытания/анализа на обнаружение нуклеиновой кислоты). Согласно описанию, имеющемуся в Стандарте валидации МЭБ, изменчивость (вариация) результатов отдельных испытаний может оцениваться в процессе множества повторяемых испытаний, включая испытания с участием двух или более операторов в течение нескольких дней. В следующих двух разделах описываются подходы к анализу данных по непрерывной и двоичной шкале с целью определения воспроизводимости испытания.

1.1. Результаты по непрерывной шкале

Самым простым подходом, необходимым для получения непрерывных результатов, является оценка среднего квадратичного отклонения (СКО) ряда образцов повторных испытаний, составляющих рабочий диапазон испытания. Изначально эти результаты должны оцениваться с помощью диаграммы расхода данных или диаграммы среднего значения повторных испытаний, отображающей зависимость от среднего квадратичного отклонения.

$$КВ = \frac{СКО\ повторных\ испытаний}{Среднее\ значение\ повторных\ испытаний}$$

где:
КВ = коэффициент вариации
СКО = среднее квадратичное отклонение

В испытаниях, в которых СКО пропорционально среднему значению повторных испытаний, часто вычисляется внутривыборочный коэффициент вариации (КВ). КВ часто используется даже в тех случаях, когда пропорциональность отсутствует. В этом случае должна сообщаться информация об уровне целевого анализируемого материала (например, низкий, средний или высокий уровень). Это необходимо, поскольку, согласно общепринятому заключению, КВ часто бывает выше в тех случаях, когда концентрация целевого анализируемого материала является низкой. Как правило, отклонение значений КВ (например, 95% доверительного интервала (ДИ)) также должно вычисляться. В тех случаях, когда значения КВ являются достаточно постоянными в ряду значений испытания, данная операция может осуществляться с использованием результатов всех образцов. В тех случаях, когда КВ является различным, исходя из значений концентрации анализируемого материала, необходимо вычислять 95% ДИ отдельно для каждой категории анализируемого материала на основании числа образцов, испытываемых на каждом уровне. Методы вычисления ДИ для КВ и разница между двумя КВ при условии использования соответствующих норме данных описаны в работе Доннера и Зоу (Donner & Zou (2012)).

Если план эксперимента включает в себя оценку множества факторов, таких как участие различных операторов и различные дни проведения эксперимента, другие подходы, такие как модели компонентов дисперсии (смешанные модели) могут быть необходимы, если целью является представить вариацию в виде суммы нескольких компонентов, которые могут быть без труда интерпретированы. Модели компонентов дисперсии также могут использоваться для получения данных о воспроизводимости (см. раздел 2).

1.2. Результаты по двоичной шкале

Как правило, количественные результаты должны использоваться для оценки точности результатов испытания в тех случаях, когда данные доступны в этой форме, хотя результаты могут быть дихотомизированы в целях составления отчета. Для испытаний с обязательным использованием двоичной шкалы, результаты которых могут быть положительными или отрицательными, для количественной оценки соглашения о результатах испытания, находящихся за пределами случайности, может использоваться каппа-статистика. Значения каппы варьируются от 0 (при отсутствии соглашения о результатах, находящихся за пределами случайности) до 1 (при наличии

абсолютного соглашения о результатах, находящихся за пределами случайности), однако в данном случае существует множество предположений, касающихся интерпретации каппы (Fleiss *et al*., 2003; Landis & Koch 1977). Наличие лучшего соглашения обычно ожидается в тех случаях, когда результаты испытаний значительно отличаются от крайних значений, и, следовательно, некоторые образцы со средними/сомнительными значениями должны испытываться во избежание чрезмерно оптимистических оценок соглашения. Взвешенная каппа для оценки результатов по порядковой шкале (например, отрицательный, сомнительный и положительный результат) может использоваться для установления того факта, что большое расхождение (например, разница в две категории) является более серьезным, чем меньшее расхождение (например, разница в одну категорию). Девяносто пять процентов ДИ должно указываться для невзвешенных или взвешенных оценок каппы (Fleiss *et al*., 2003).

Таблица 1. *Примеры вычисления каппы для результатов по двоичной шкале*

Пример 1: Вычисление каппы, основанное на результатах повторных испытаний, классифицируемых как положительные или отрицательные

Результат испытания	***Положительный результат***	***Отрицательный результат***
Положительный	90	5
Отрицательный	10	95
	100	100

Каппа = 0,85 (95% ДИ = от 0,78 до 0,92)

Пример 2: Вычисление каппы, основанное на результатах повторных испытаний, делящихся на три группы (положительные, сомнительные и отрицательные результаты)

Результат испытания	***Положительный результат***	***Сомнительный результат***	***Отрицательный результат***
Положительный	80	10	10
Сомнительный	15	75	10
Отрицательный	5	15	80
	100	100	100

Каппа = 0,68 (95% ДИ = от 0,61 до 0,75). Взвешенная каппа = 0,70. (95% ДИ = от 0,61 до 0,79)

2. Воспроизводимость испытания в нескольких лабораториях

Точность результатов испытания варьируется в зависимости от условий проведения испытания, например, в зависимости от наличия различных операторов, различных мест проведения испытания, использования различных комплектов для проведения испытания или от проведения испытания в разные дни. Чаще всего термин *воспроизводимость* применяется к оценке точности выбранного испытания, проводимого в нескольких лабораториях. Коэффициенты, остающиеся неизменными, должны описываться с той целью, чтобы дать возможность интерпретировать результаты в контексте фактической ситуации. Исследования воспроизводимости могут проводиться независимо от исследований повторяемости или в сочетании с исследованиями повторяемости, однако проводиться они должны в слепом режиме. Согласно рекомендациям Стандарта валидации МЭБ, как минимум три лаборатории должны провести испытания как минимум 20 образцов с идентичными аликвотами, предоставленными каждой лаборатории.

Статистические методы анализа исследований воспроизводимости испытания в нескольких лабораториях являются идентичными методам, используемым для оценки повторяемости испытаний в лаборатории. Несмотря на это, в качестве части межлабораторного исследования важным может быть оценка и классификация вариации в результатах исследования, полученных из нескольких источников (часто именуемая «классом»).

Внутриклассовый (внутригрупповой) коэффициент корреляции (ВКК) представляет собой сходство или корреляцию любых двух измерений, выполненных на одном и том же образце. Значение ВКК составляет от 0 до 1. При этом значения, приближающиеся к 1, указывают на наличие минимальной погрешности измерения. Напротив, значения, приближающиеся к 0, указывают на наличие значительной погрешности измерения.

Например, исследование планировалось с целью проверки результатов испытания в трех лабораториях, в каждой из которых работу выполняет два высококвалифицированных специалиста и используются образцы для повторного испытания в двух сериях комплектов для испытаний. Каждый образец для испытаний будет испытываться 24 раза. Отобранные факторы испытания (лаборатория, специалисты, серия комплектов для испытаний, результат повторных испытаний) могут рассматриваться как постоянные или как случайные факторы, в зависимости от того, как они отбирались и являются ли они характерными для целевой популяции. При выполнении данного плана исследования компоненты вариации могут оцениваться по каждому классу (пример: Dargatz *et al.,* 2004), а внутриклассовый (внутригрупповой) коэффициент корреляции (ВКК) может оцениваться в качестве меры схожести выборочных результатов (Bartlett & Frost, 2008).

2.1. Техническая модификация метода исследования

После осуществления валидации испытания для использования в контролируемой лабораторной среде может быть рассмотрена возможность его использования в значительно отличающейся среде (такой, как среда для выполнения теста для экспресс-диагностики на месте). По причине более существенных изменений, например, резких перепадов температур, которые часто наблюдаются при выполнении теста для экспресс-диагностики на месте, можно спрогнозировать, что два испытания будут протекать по-разному в своих различающихся средах. Фактически, в значительно большей степени, чем случайная погрешность измерения, применяемая к оценке внутрилабораторных и межлабораторных погрешностей измерений, предполагается, что значения такого исследования с большей вероятностью будут интерпретироваться как систематическая погрешность измерений, которая будет возникать, если данные значения способствуют переоценке и недооценке действительного значения. Например, при проведении теста для экспресс-диагностики на месте (в полевых условиях) и испытания (теста) в условиях лаборатории, среднее разниц между значением, полученным в результате теста для экспресс-диагностики на месте, и значением, полученным в результате испытания в лаборатории (действительным значением) для одного и того же образца должно указываться с 95% ДИ. Если в 95% ДИ не входит ноль, это является свидетельством систематической погрешности результатов теста для экспресс-диагностики на месте в сравнении с испытанием в лаборатории. В случае наличия систематической погрешности в результатах теста, результаты теста для экспресс-диагностики на месте не подлежат сравнению с результатами лабораторного валидированого анализа. Для валидации теста для экспресс-диагностики на месте тест либо подвергается «технической модификации», после чего оценивается с помощью сравнительного исследования методов (см. главу 3.6.8), либо требуется полная ревалидация теста для экспресс-диагностики на месте.

Аналогические подходы могут использоваться для лабораторной оценки изменений в методах с целью определения систематической или случайной вариации в результатах.

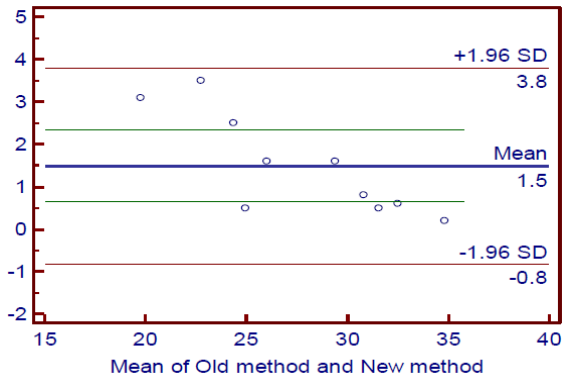
Пример: Следующие неопубликованные данные были получены путем сравнения двух методов экстракции (старого и нового, применяемых к разделенным образцам) при значениях порога цикла (ПЦ) для количественной полимеразной цепной реакции (кПЦР) в масштабе реального времени,

соответствующих катаральной лихорадке овец. Данные (n=10) представляют собой средние значения для образцов для повторных испытаний.

Старый метод: 25,6, 24,5, 21,3, 26,8, 25,2, 30,2, 31,2, 32,8, 31,8, 34,9

Новый метод: 23,1, 21,0, 18,2, 25,2, 24,7, 28,6, 30,4, 32,2, 31,3, 34,7

Средняя разница между двумя методами (старый минус новый) составила –1,49 (95% ДИ = от –2,33 до – 0,64) при двусторонней вероятности p=0,003. Поскольку 95% ДИ не включает в себя ноль, это свидетельствует о систематическом более низком значении ПЦ при использовании нового метода экстракции. График Бланда-Альтмана (Bland & Altman, 1999; Fig. 3) может использоваться для графического отображения изменения разницы как функции среднего значения старого и нового метода. В этих данных наблюдается снижение разницы для более высоких значений ПЦ, но размер образца является маленьким.



SD	СКО
Mean	Среднее значение
Mean of Old method and New method	Среднее значение старого и нового методов

Рисунок 3. *График Бланда-Альтмана для средней разницы (ось y) в значениях ПЦ как функции среднего значения старого и нового методов (n=10).*

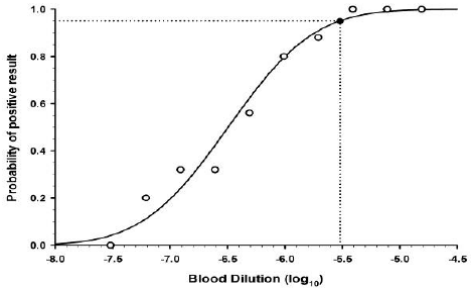
3. Аналитическая чувствительность (АЧ, синоним = нижний предел чувствительности: НПЧ)

Аналитическая чувствительность может оцениваться с помощью эксперимента «разведение к уменьшению» (ЭРУ), в котором серийное разведение известного установленного количества целевого анализируемого материала осуществляется в соответствующей матрице проб (образцов). Это известное установленное количество может быть взято из местного или национального/международного эталонного стандарта или рабочей пробы, концентрация аналитического материала в которых определена. Параллельное использование эталона сравнения возможно, но не является обязательным, если проводимое исследование не является таким, в котором минимальное изменение в валидированном анализе сравнивается с изначальным состоянием валидированного анализа. Подход ЭРУ может использоваться в тех случаях, когда количество анализируемого материала определяется качественным или количественным путем. В последнем случае результат испытания подлежит повторной классификации как положительный или отрицательный.

Подход к анализу данных НПЧ зависит от плана эксперимента. Например, предположим, что было проведено исследование, в котором 108 колониеобразующих единиц (КОЕ) бактерии было добавлено в 10 г экскрементов, и была достигнута концентрация 107 КОЕ/г. Затем этот образец был разбавлен с помощью десяти последовательных разбавлений, после которых концентрация достигла 101 КОЕ/г. Эксперимент был проведен повторно трижды. Если все образцы повторных исследований, обнаруживаемые при значении 103 КОЕ/г, отсутствовали при значении 102 КОЕ/г,

НПЧ в таком случае, согласно осторожным предварительным подсчетам, составлял 103 КОЕ/г. Если было необходимо определить точное значение НПЧ, планировался второй этап эксперимента по определению НПЧ с большей точностью с использованием ряда более тонких разведений, например, двукратных разведений, охватывающих промежуток между значением 100% обнаружения и значением 0% обнаружения, установленный во время первого эксперимента. Часто в качестве предельного значения НПЧ выбирается 95%. В эксперименте с 20 образцами для повторных испытаний это значение соответствует разведению, при котором 19 образцов анализируемого материала для повторного разведения было положительным. Важным пунктом является то, что выбранное значение вероятности НПЧ (95% или 50% другого значения) должно определяться и использоваться последовательно в случае сравнения результатов многократных испытаний. НПЧ может оцениваться с помощью непараметрического подхода Спирмена-Кербера, а также с помощью логической регрессии или пробит-анализа. Чем больше число повторов для каждого разведения, тем более точной является оценка НПЧ.

Пример: Была выполнена серия двукратных разведений крови лошади, показавшая положительный результат в тесте на наличие вируса африканской чумы лошадей (ВАЧЛ) (10^{-3} разведений), охватывавшую нелинейный диапазон испытания (Guthrie *et al*., 2013). Экстракция была повторно проведена 25 раз, и образцы были испытаны на наличие ВАЧЛ методом кПЦР в масштабе реального времени. Результаты кПЦР для 15 разведений были использованы в пробит-анализе для вычисления 95% НПЧ (т. е. исходной концентрации, дающий положительный результат при проведении кПЦР в масштабе реального времени в 95% повторов (Burns & Valdivia, 2008). В результате испытания было установлено, что 95% НПЧ отмечает значение разведения, равное $3{,}02 \times 10^{-6}$, как показано на рисунке 4, что соответствует значению цикла количественного анализа 35,71 при проведении кПЦР. Значение ДИ, использовавшееся при оценке, указано не было.



Probability of positive result	Вероятность положительного результата
Blood Dilution ($\log_{10}$)	Разведение крови (десятичный логарифм)

Рисунок 4. *Установлено значение 95% нижнего предела чувствительности для ВАЧЛ в крови лошади (десятичный логарифм), показанное пунктирной линией*

4. Аналитическая специфичность (АС)

Аналитическая специфичность может быть описана как минимум тремя различными методами: избирательности, эксклюзивности (синоним: кривая зависимости перекрестной реакции) и инклюзивности (в соответствии с описанием, приведенным в Стандарте валидации МЭБ). Значения, полученные двумя последними методами измерения, должны указываться на основе показателей клеточной линии, изолятов, видов или родов, соответствующих целевому анализируемому материалу и целевому назначению испытания. При проведении скрининг-тестов необходим более широкий диапазон и более высокий показатель иклюзивности специфичности в сравнении с контрольным тестом (испытанием), который может разграничивать изоляты, различающиеся, например, по своей патогенности. Поскольку выбор связанных организмов является субъективным и часто зависит от типа и числа образцов, результат эксклюзивности должен указываться, исходя из качественной оценки, например, процент связанных агентов, демонстрирующих перекрестную реакцию при проведении анализа с использованием перечня

потенциальных агентов, демонстрирующих перекрестную реакцию, которые были оценены. Аналогичным образом, инклюзивность отображается как процент серологических вариантов, штаммов, родов и видов, обнаруженных в результате исследования и соответствующих целевому анализируемому материалу.

5. Диагностическая эффективность анализа

Диагностическая эффективность анализа наиболее часто измеряется как диагностическая чувствительность (ДЧ) или диагностическая специфичность (ДС) или объединенное измерение ДЧ и ДС, такое как коэффициент вероятности положительного или отрицательного результата. Коэффициенты вероятности для различных интервалов результатов анализа также могут вычисляться в тех случаях, когда более важно зафиксировать информацию о величине (диапазоне) результатов испытания, чем использовать ее в дихотомизированной форме. Существуют публикации, где указана более подробная информация об использовании и вычислении коэффициентов вероятности (Gardner & Greiner, 2006; Gardner *et al.*, 2010). Последняя работа включает в себя пример вычисления ДИ двумя способами в случаях диагностики токсоплазмоза свиней.

ДЧ и ДС может оцениваться в случаях, когда эталонный или сравнительный метод является абсолютно чувствительным и специфичным или когда эталонный стандарт является несовершенным. Как правило, большинство эталонных стандартов предубойных исследований, повсеместно используемых в диагностических лабораториях, являются несовершенными, в связи с чем часто возникает необходимость проведения некропсии с исследованием множества образцов тканей с помощью дополнительных тестов и испытаний, таких как культуральное и/или гистопатологическое исследование, если результаты эталонного стандарта считать действительными. Для большинства исследований с целью валидации тестов и испытаний, предназначенных для диагностики болезней у животных, последний пункт не является целесообразным или экономически эффективным за исключением ограниченного числа образцов.

5.1. ДЧ и ДС при использовании эталонного стандарта

Потенциальное испытание может давать результаты по двоичной (дихотомической), порядковой (например, титр) и непрерывной шкалам. При использовании последних двух шкал результаты должны быть дихотомизированы перед вычислением ДЧ и ДС, т. е. необходимо определить критическое (пороговое) значение. Для определения ДЧ и ДС рекомендуется использовать точные биноминальные значения 95% ДИ (Greiner & Gardner, 2000), поскольку нормальная аппроксимация может не давать соответствующего значения ДИ в тех случаях, когда значение параметра оценивается как близкое к 1.

Статистическая погрешность при определении параметров диагностической эффективности, например ДЧ и ДС, должна быть представлена в виде доверительного интервала (ДИ). Как правило, используется значение 95% ДИ, и ширина интервала (точность оцененного значения) в значительной степени зависит от размера образца, используемого для оценки параметра. Использование точных значений ДИ является более предпочтительным, чем использование нормальных аппроксимаций, поскольку последние не охватывают верхние предельные значения, превышающие 100%.

Пример: непрямой твердофазный иммуноферментный анализ (непрямой твердофазный ИФА)

	Число животных	
	Положительный результат испытания на наличие	Отрицательный результат испытания на наличие

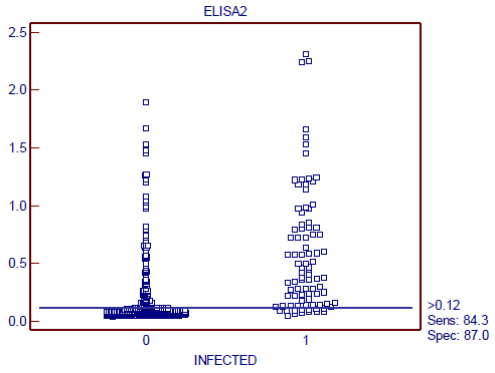
		известного антитела (369)		известного антитела (198)	
Результаты испытания	Положительный	287	ИПР	ЛПР	1
	Отрицательный	82	ЛОР	ИОР	197

	Диагностическая чувствительность* ИПР/(ИПР + ЛОР)	Диагностическая специфичность* ИОР/(ИОР + ЛПР)
	77,8% (73,2 – 81,9%)*	**99,5%** (97,2 – 99,9%)*

ИПР и ЛПР = истинно положительный результат и ложноположительный результат, соответственно
ИОР и ЛОР = истинно отрицательный результат и ложноотрицательный результат, соответственно
* 95% – точные биноминальные значения доверительных интервалов для ДЧ и ДС

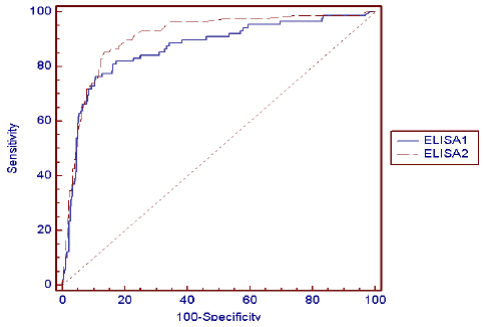
Когда эталонный стандарт не применяется ко всем положительным и отрицательным результатам испытания (частичная верификация), то для учета различных производящих вероятностей в группах с положительным и отрицательным результатами испытаний должна осуществляться исправленная оценка ДЧ и ДС (Greiner & Gardner, 2000).

Для анализов, демонстрирующих результаты по порядковой шкале (например, значения титра) или непрерывной шкале (например, коэффициенты образцов испытаний для определения значений положительных контрольных образцов в твердофазном ИФА), оценки ДЧ и ДС должны дополняться оценками области ниже кривой операционной характеристики приемника (кривой ОХП). Анализ ОХП обеспечивает подход, независимый от критического значения, необходимый для оценки общей точности испытания, результаты которого отображаются в виде порядковых или непрерывных значений. Область ниже кривой ОПХ представляет собой одинарную числовую оценку общей точности, варьирующуюся от 0,5 (непригодное испытание) до 1 (испытание, отвечающее всем требованиям). Главным подтверждением результатов анализа ОПХ является то, что критические значения, используемые для интерпретации результатов испытания, могут варьироваться в зависимости от цели проведения испытания (например, скрининг в сравнении с подтверждением), и с распространенностью инфекции, от расходов на ошибки испытания, а также от наличия других испытаний. Подробные описания анализа ОПХ представлены и в других работах (Gardner & Greiner, 2006; Greiner *et al*., 2000; Zweig & Campbell, 1993). В случае сравнения результатов многократных испытаний с использованием порядковой или непрерывной шкалы необходимо вычислить разницу в области ниже кривой при 95% ДИ. Методы вычисления разницы различаются для независимых и зависимых образцов и применяются во многих статистических программах (Gardner & Greiner, 2006). Примеры точечного графика, отображающего результаты однократного твердофазного ИФА, и кривых ОПХ для двух твердофазных ИФА представлены на рисунках 5 и 6.



ELISA2	Твердофазный ИФА 2
INFECTED	ИНФИЦИРОВАННЫЕ ЖИВОТНЫЕ
Sens.	Чувствительность
Spec.	Специфичность

Рисунок 5. *Точечный график результатов твердофазного ИФА для неинфицированных (код = 0) и инфицированных (код = 1) животных*



Sensitivity	Чувствительность
ELISA1	Твердофазный ИФА 1
ELISA2	Твердофазный ИФА 2
Specificity	Специфичность

Рисунок 6. *Кривая операционной характеристики приемника для двух твердофазных ИФА*

При отсутствии отвечающего всем требованиям эталонного стандарта существует также возможность оценки ОНК с использованием моделей латентных классов (ЛК). Например, модели ЛК могут применяться к нормально распределенным данным двух зависимых испытаний (Choi *et al*., 2003) и при использовании полупараметрических подходов (Branscum *et al*., 2008). Модели ЛК для непрерывных данных, включающие цензурированные или усеченные данные, возникающие при проведении испытаний методом ПЦР в масштабе реального времени, не описаны в настоящей инструкции по валидации по причине их сложности. Несмотря на это, модели ЛК для результатов испытаний с использованием двоичной шкалы, а также пример их использования описаны в разделе 5.3.

5.2. Сравнение оценок ДЧ и ДС для двух испытаний с использованием отвечающего всем требованиям эталонного стандарта

Очень часто исследователи хотят сравнить значения ДЧ в субпопуляциях инфицированных животных, например, значения для инфицированных животных с клиническими симптомами со значениями для инфицированных животных с субклиническими симптомами, или значения ДЧ в различных географических областях. Поскольку данные образцы являются независимыми, сравнения могут выполняться статистически с использованием критерия хи-квадрат Пирсона для обеспечения однородности. В качестве альтернативы могут быть раздельно вычислены значения 95% ДИ и 95% ДИ разницы в двух пропорциях. В случае сравнения ДЧ (или ДС) двух испытаний на одном и том же наборе инфицированных (или неинфицированных) образцов сдвоенным методом результаты испытаний более не являются независимыми. Статистические методы, такие как критерий хи-квадрат МакНемара, могут использоваться для определения гипотетических равных значений чувствительности (специфичности) при проведении испытаний на одних и тех же образцах.

Пример: Пять испытаний на обнаружение антител были оценены с целью диагностирования паратуберкулеза коров у дойных коров в известных инфицированных и неинфицированных стадах, наличие которого было установлено в результате микроскопии экскрементов и истории болезней стада. Следующие таблицы с данными были составлены на основе исходных данных перед их

последующей публикацией (Collins *et al*., 2005). В этой публикации из результатов анализа было исключено одно стадо. Данный пример используется для демонстрации целей, связанных с отображением в форме таблицы плана вычисления ДЧ и ДС, а также со статистической оценкой.

Инфицированные животные

	T_2^+	T_2^-	
T_1^+	124	74	198
T_1^-	8	243	251
	132	317	449

Неинфицированные животные

	T_2^+	T_2^-	
T_1^+	3	27	30
T_1^-	16	366	382
	19	393	412

Чувствительность T1 = 198/449 = 44,1%

Чувствительность T2 =132/449 = 29,4%

Специфичность T1 = 382/412 = 92,7%

Специфичность T2 = 393/412 = 95,4%

Значения чувствительности значительно разнятся (p<0,0001), но в случае со значениями специфичности такого не наблюдается (p= 0,126). Значения вычислены на основе данных двустороннего критерия хи-квадрат МакНемара. Ковариантны чувствительности и специфичности (см. Gardner *et al*., 2000 for details) также могут вычисляться для определения случаев, в которых результаты испытаний являются условно независимыми или зависимыми, что позволяет определить статус инфекции. Для этих данных ковариантa чувствительности (вычисленная с помощью таблицы с результатами инфицированных животных слева) составила 0,147 (p < 0,0001 в соответствии с результатами критерия хи-квадрат Пирсона), что является показателем сильной зависимости двух результатов испытаний на инфицированных животных. Ковариантa специфичности (вычисленная с помощью таблицы с результатами неинфицированных животных справа) составила 0,004 (p = 0,152 в соответствии с результатами критерия хи-квадрат Пирсона), что является показателем незначительной зависимости.

Существует публикация, в которой представлен дополнительный пример, основанный на данных диагностики токсоплазмоза свиней (Gardner *et al*., 2010).

5.3. Определение ДЧ и ДС без использования отвечающего всем требованиям эталонного стандарта

Развитие статистической методологии, в частности разработка моделей латентных классов (иногда называемых «классами, не относящиеся к золотому стандарту»), в настоящее время позволяет исследователям освободить себя от использования ограничительного предположения, связанного с проведением отвечающего всем требованиям стандартного контрольного испытания, оценкой точности потенциального(ых) испытания(ий) и с эталонным стандартом, использующим те же данные (Enoe *et al*., 2000; Hui & Walter, 1980).

Модели латентных классов (ЛК), во всех из которых используется метод максимального правдоподобия или байесовский метод, могут использоваться для оценки ДЧ и ДС в тех случаях, когда доступны объединенные результаты различных испытаний, применявшихся к животным в различных популяциях (например, в стадах или географических зонах). Не все модели ЛК для оценки ДЧ и ДС будут поддаваться статистической идентификации с целью составления заключения. Модель является идентифицируемой, если теоретически возможно определить истинное значение параметров модели после проведения неограниченного числа исследований такой модели. В сущности, это можно отождествить с наличием уникального ряда значений представляющих интерес параметров (ДЧ и ДС). Байесовские методы особенно подходят для использования в тех случаях, когда имеется предварительная информация о ДЧ и/или ДС и когда проблема оценки не является идентифицируемой (Branscum *et al*., 2005).

Самой простой моделью ЛК, основанной на одной популяции, является модель, идентифицируемая в тех случаях, когда три условно независимых испытания проводятся на одних и тех же образцах. Ограничение независимости трех испытаний может быть сложной задачей на

практике, если целевой анализируемый материал не различается в различных испытаниях. Исходя из этого, повсеместно используемым в ветеринарной практике подходом является проведение двух испытаний на всех образцах, полученных у животных в двух популяциях, поскольку такой метод является менее затратным, и предположения об условной независимости могут быть более обоснованными. Модель, основанная на проведении двух испытаний в двух популяциях, также требует предположения о постоянной чувствительности и специфичности в двух популяциях, а также наличия четких показателей распространения. Подтверждение предположения о постоянной чувствительности может быть сложной процедурой, а корректность самого такого предположения может быть маловероятной, если в одной популяции имеются зараженные животные с клиническими симптомами, а в другой популяции имеются зараженные животные с субклиническими симптомами, поскольку множество опубликованных исследований показало, что чувствительность испытания является более высокой для зараженных животных с клиническими симптомами. Если известно, что у животных одной из двух популяций отсутствует патоген (распространенность равна нулю), в то время как у животных другой популяции распространенность не равна нулю, первая популяция может быть использована для оценки ДС, что упростит оценку ДЧ в инфицированной популяции.

Перечень болезней МЭБ, оценка ДЧ и ДС методов идентификации которых производилась с использованием байесовских методов, включает в себя бруцеллез овец (Praud *et al*., 2012), австралийскую лихорадку Q (Paul *et al*., 2013), трипанозомоз (Bronsvoort *et al*., 2010), туберкулез крупного рогатого скота (Clegg *et al*., 2011), ящур (Bronsvoort *et al*., 2006), африканскую чуму лошадей (Guthrie *et al*., 2011) и инфекционную анемию лососевых (Caraguel *et al*., 2013).

Программа WinBUGS[1] позволяет легко использовать анализ Монте-Карло с использованием цепей Маркова в оценке байесовским методом (Lunn *et al.,* 2000) и упрощает анализ максимальной вероятности, который может проводиться с использованием веб-интерфейса (Poulliot *et al.,* 2002). Предварительная информация о параметрах модели, используемая в байесовском анализе, может оказывать влияние на окончательные результаты оценки, в зависимости от относительной силы свидетельств, обеспечиваемой априорной вероятностью (уровень априорной неопределенности) и данными (неопределенность, связанная с конечными размерами образцов). Таким образом, источники предварительной информации должны быть надлежащим образом зафиксированы в описаниях байесовских анализов. Также желательно выполнить повторный анализ с использованием неинформативных значений априорной вероятности для всех параметров в том случае, когда модель является идентифицируемой.

Максимальная вероятность – метод оценки наиболее вероятных значений представляющих интерес параметров, основанный на функции вероятности данных.

Байесовские методы – включают в себя соответствующую предварительную информацию или предварительное знание об одном или более испытаниях в дополнение к функции вероятности данных. При использовании образцов большого размера, максимальной вероятности и байесовских методов выводы являются схожими.

Важно отметить, что при выполнении анализа ЛК не может делаться поправка на погрешности, характерные для исследований с ненадлежащим образом разработанным планом. Методы должны использоваться аккуратно и включать в себя полную оценку соответствующих предположений (например, предположений об условной зависимости, постоянной чувствительности и специфичности во всех популяциях, а также предположений о распространенности, носящей отчетливый характер), влияние использования выбранных показателей априорного распределения на выводы апостериорного распределения, как описано в предыдущем параграфе, а также конвергенцию цепей Маркова в анализе, проводимом байесовским методом (Toft *et al.,* 2005).

Пример: Была проведена оценка ДЧ и ДС метода количественной ПЦР в масштабе реального времени и метода условного выделения вируса (ВВ) для обнаружения вируса африканской чумы

[1] Программа доступна по адресу: http://www.mrc-bsu.cam.ac.uk/bugs/winbugs/contents.shtml

лошадей (АЧЛ) в образцах цельной крови с помощью байесовской модели латентных классов при проведении двух испытаний в двух популяциях (Guthrie *et al*., 2013). Две популяции южноафриканских чистокровных лошадей (503 особи с подозрением на наличие вируса АЧЛ и 503 здоровые лошади из зоны контроля распространения вируса АЧЛ) были исследованы методами ПЦР и ВВ. Объединенные результаты по 503 особям с подозрением на наличие вируса были следующими: ПЦР+ВВ+ (*n*=156). ПЦР+ВВ– (*n*=184), ПЦР–ВВ+ (*n*=0) и ПЦР–ВВ– (*n*=163). Все 503 здоровые лошади продемонстрировали результат ПЦР–ВВ–. Различные модели (условной независимости и условной зависимости) были согласованы с данными, и вторая популяция здоровых лошадей была также включена в некоторые анализы.

Модели были созданы в программе WinBUGS 1.4.3 (Lunn *et al.,* 2000) с отказом от первых 5 000 рабочих циклов и использованием следующих 50 000 рабочих циклов в выводах апостериорного распределения (срединные значения и 95% интервалы вероятности для ДЧ и ДС. Конвергенция модели оценивалась путем визуальной проверки трассировочного графика повторяющихся значений и существующих кратных цепей из распределенных исходных значений. Модель условной независимости, согласованная с неинформативными бета-значениями (1,1) априорной вероятности, касающимися ДЧ и ДС обоих испытаний, показала результаты, практически идентичные результатам модели, в которой использовались высокоинформативные бета-значения (9999,1) априорной вероятности, касающиеся ДЧ метода ВВ. Оцененные срединные значения и 95% интервалы вероятности (иногда называемые доверительными интервалами), указанные в круглых скобках, из модели условной независимости с неинформативными значениями априорной вероятности, были следующими:

Чувствительность ПЦР = 0,996 (0,977–0,999)

Специфичность ПЦР = 0,999 (0,993–1,0)

Чувствительность ВВ = 0,458 (0,404–0,51)

Специфичность ВВ = 0,999 (0,998–1,0)

Полученные результаты показали, что ДЧ ПЦР вдвое выше ДЧ ВВ, а ДС обоих испытаний является сопоставимым. Существует публикация с полным описанием модельного подхода (Guthrie *et al.*, 2013).

5.4. Сравнение оценок ДЧ и ДС двух испытаний без использования отвечающего всем требованиям эталонного стандарта

Если в программе WinBUGS байесовский подход используется с целью анализа объединенных данных испытаний, проводившихся в различных популяциях, разница в чувствительности (специфичности) может быть легко оценена, а вероятность того, что чувствительность (специфичность) одного из исследований выше чувствительности (специфичности) другого исследования, может быть оценена с помощью единичной ступенчатой функции (функции Хевисайда).

Пример: В результатах, изложенных в разделе 5.3, 95% интервалы вероятности (ИВ) для ДЧ не перекрывались, однако заметное перекрытие наблюдалось в 95% ИВ для ДС (Guthrie *et al.,* 2013). Соответствующие значения вероятности, полученные с помощью единичной ступенчатой функции (функции Хевисайда), составили 1 и 0,24, соответственно. Эти значения указывают на несомненность того факта, что ДЧ методов различается, однако вероятность того, что ДС также различается, является низкой (ниже 0,5).

СПИСОК ЛИТЕРАТУРЫ

BARTLETT J.W. & FROST C. (2008). Reliability, repeatability and reproducibility: analysis of measurement errors in continuous variables. *Ultrasound Obstet. Gynecol.*, **31**, 466–475.

BLAND J.M. & ALTMAN D.G. (1999). Measuring agreement in method comparison studies. *Statist. Methods Med. Res*., **8**, 135–160.

BOSSUYT P.M., REITSMA J.B., BRUNS D.E., GATSONIS C.A., GLASZIOU P.P., IRWIG L.M., LIJMER J.G., MOHER D., RENNIE D. & H.C.M. DE VET (2003). Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. *Clin. Chem.*, **49**, 1–6.

BRANSCUM A.J., GARDNER I.A. & JOHNSON W.O. (2005). Estimation of diagnostic-test sensitivity and specificity through Bayesian modeling. *Prev. Vet. Med*., **68**, 145–163.

BRANSCUM A.J., JOHNSON W.O., HANSON T.E. & GARDNER I.A. (2008). Bayesian semiparametric ROC curve estimation and disease diagnosis. *Stat. Med*., **17**, 2474–2496.

BRONSVOORT B.M., TOFT N., BERGMANN I.E., SØRENSEN K.J., ANDERSON J., MALIRAT V., TANYA V.N., MORGAN K.L. (2006) Evaluation of three 3ABC ELISAs for foot-and-mouth disease non-structural antibodies using latent class analysis. *BMC Vet. Res*., **2**, 30.

BRONSVOORT B.M., VON WISSMANN B., FÈVRE E.M., HANDEL I.G., PICOZZI K., & WELBURN S.C. (2010) No gold standard estimation of the sensitivity and specificity of two molecular diagnostic protocols for *Trypanosoma brucei* spp. in Western Kenya. *PLoS One*; **5** (1), e8628.

BURNS M & VALDIVIA H. (2008). Modelling the limit of detection in real-time quantitative PCR. *Eur. Food Res. Technol.*, **226**, 1513–1524.

CARAGUEL C., STRYHN H., GAGNE N., DOHOO I. & HAMMELL L. (2012). Use of a third class in latent class modelling for the diagnostic evaluation of five infectious salmon anaemia virus detection tests. *Prev. Vet. Med.*, **104**, 165–173.

CHOI Y.K., JOHNSON W.O., COLLINS M.T. & GARDNER I.A. (2006). Bayesian inferences for receiver operating characteristic curves in the absence of a gold standard. *J. Agric, Biol. Environ. Stat.*, **11**, 201–229.

CLEGG T.A., DUIGNAN A., WHELAN C., GORMLEY E., GOOD M., CLARKE J., TOFT N. & MORE S.J. (2011). Using latent class analysis to estimate the test characteristics of the γ-interferon test, the single intradermal comparative tuberculin test and a multiplex immunoassay under Irish conditions. *Vet. Microbiol*.,**151**, 68–76.

COLLINS M.T., WELLS S.J., PETRINI K.R., COLLINS J.E., SCHULTZ R.D., & WHITLOCK R.H. (2005). Evaluation of five antibody detection tests for diagnosis of bovine paratuberculosis. *Clin. Diag. Lab. Immunol*., **12**, 685-692.

DARGATZ D.A., BYRUM B.A., COLLINS M.T., GOYAL S.M., HIETALA S.K., JACOBSON R.H., KOPRAL C.A., MARTIN B.M., MCCLUSKEY B.J. & TEWARI D. (2004). A multilaboratory evaluation of a commercial enzyme-linked immunosorbent assay test for the detection of antibodies against *Mycobacterium avium* subsp. *paratuberculosis* in cattle. *J. Vet. Diagn. Invest.*, **16** (6), 509–514.

DONNER A & ZOU G.Y. (2012). Closed-form confidence intervals for functions of the normal mean and standard deviation. *Stat. Meth. Med. Res*., **21** (4), 347–359.

ENØE C., GEORGIADIS M.P. & JOHNSON W.O. (2000). Estimation of sensitivity and specificity of diagnostic tests and disease prevalence when the true disease state is unknown. *Prev. Vet. Med.*, **45**, 61–81.

FLEISS J.L., LEVIN B. & PAIK M.C. (2003). Statistical Methods for Rates and Proportions, Third Editionn. John Wiley & Sons, New York, USA.

GARDNER I.A., STRYHN H., LIND P., & COLLINS M.T. (2000). Conditional dependence between tests affects the diagnosis and surveillance of animal diseases. *Prev. Vet. Med.,* **45**, 107–122.

GARDNER I.A. & GREINER M. (2006). Receiver-operating characteristic curves and likelihood ratios: improvements over traditional methods for the evaluation and application of veterinary clinical pathology tests. *Vet. Clin. Pathol.*, **35**, 8–17.

GARDNER I.A., GREINER M. & DUBEY J.P. (2010). Statistical evaluation of test accuracy studies for *Toxoplasma gondii* in food animal intermediate hosts. *Zoonoses Public Health*, **57**, 82–94.

GREINER M. & GARDNER I.A. (2000). Epidemiologic issues in the validation of veterinary diagnostic tests. *Prev. Vet. Med.,* **45**, 3–22.

GREINER M., PFEIFFER D. & SMITH R.D. (2000). Principles and practical application of the receiver-operating characteristic analysis for diagnostic tests. *Prev. Vet. Med.,* **45**, 23–41.

GUTHRIE A.J., MACLACHLAN N.J., JOONE C., LOURENS C.W., WEYER C.T., QUAN M., MONYAI M.S. & GARDNER I.A. (2013). Diagnostic accuracy of a duplex real-time reverse transcription quantitative PCR assay for detect ion of African horse sickness virus. *J. Virol. Methods,* **189**, 30–35.

HUI S.L. & WALTER S.D. (1980). Estimating the error rates of diagnostic tests. *Biometrics*, **36**, 167–171.

LANDIS J.R. & KOCH G.G. (1977). The measurement of observer agreement for categorical data. *Biometrics,* **33**, 159–174.

LUNN D.J., THOMAS A., BEST N. & SPIEGELHALTER D. (2000). WinBUGS – a Bayesian modelling framework: concepts, structure, and extensibility. *Statist. Comp.,* **10**, 325–337.

PAUL S., TOFT N., AGERHOLM J.S., CHRISTOFFERSEN A.B. & AGGER J.F. (2013). Bayesian estimation of sensitivity and specificity of *Coxiella burnetii* antibody ELISAs in bovine blood and milk. *Prev. Vet. Med.,* **109**, 258–263.

PRAUD A., CHAMPION J.L., CORDE Y., DRAPEAU A., MEYER L. & GARIN-BASTUJI B. (2012) Assessment of the diagnostic sensitivity and specificity of an indirect ELISA kit for the diagnosis of *Brucella ovis* infection in rams. *BMC Vet. Res.,* **8**, 68.

POUILLOT R., GERBIER G. & GARDNER I.A. (2002). "TAGS", a program for the evaluation of test accuracy in the absence of a gold standard. *Prev. Vet. Med.,* **53**, 67–81.

TOFT N., JORGENSEN E. & HOJSGAARD S. (2005). Diagnosing diagnostic tests: evaluating the assumptions underlying the estimated of sensitivity and specificity in the absence of a gold standard. *Prev. Vet. Med.,* **68**, 19–33.

WILKS C. (2001). Gold standards as fool's gold. *Aust. Vet. J.*, **79**, 115.

ZWEIG M.H. & CAMPBELL G. (1993). Receiver-operating characteristic (ROC) plots - a fundamental evaluation tool in clinical medicine. *Clin. Chem.*, **39**, 561–577.

* * *