CHAPITRE 2.2.5.

METHODES STATISTIQUES DE VALIDATION

INTRODUCTION

Les Recommandations de l'OMSA pour la validation fournissent des informations détaillées et des exemples à l'appui de la Norme de validation publiée au Chapitre 1.1.6 Validation des épreuves diagnostiques pour les maladies infectieuses des animaux terrestres *du présent* Manuel terrestre*. L'expression « Norme de validation de l'OMSA » dans le présent chapitre doit être comprise comme renvoyant à ce chapitre.*

Le choix des méthodes statistiques pour l'analyse des données de validation des tests, générées lors d'expériences en laboratoire ou d'évaluations d'échantillons de terrain, dépend de considérations comme le modèle expérimental ou la sélection des échantillons (provenance, nombre d'échantillons, nombres de répétitions du test, etc.). Le choix de la « meilleure méthode » doit être fait de concert avec un statisticien durant la phase de conception avant que les études de validation ne débutent.

Définition des échelles de mesure

Binaire (dichotomique): soit positif, soit négatif parce qu'il s'agit de la manière dont les résultats du test sont présentés ou positif/négatif à la valeur seuil choisie lorsque les résultats sont mesurés sur une échelle ordinale ou continue.

Ordinale: mesuré sur une échelle avec des valeurs discrètes ; des valeurs plus élevées indiquent généralement plus d'analyte, p. ex. titres de séroneutralisation virale.

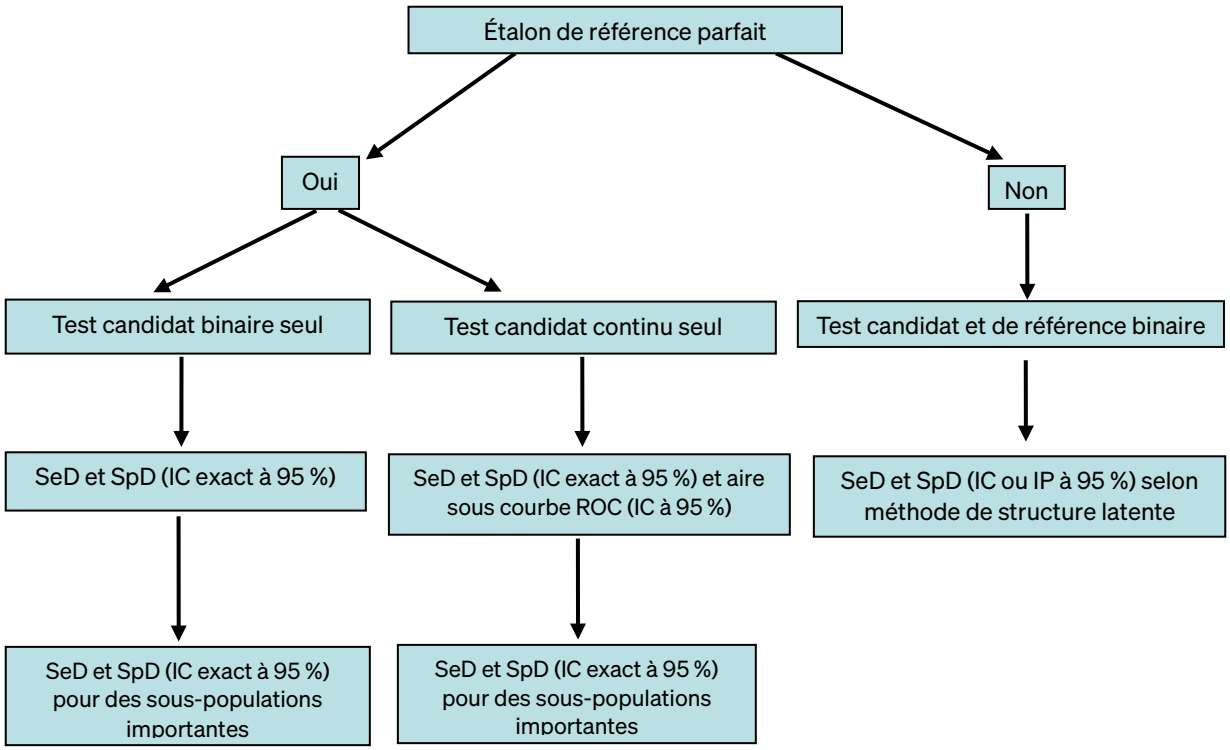
Continue: un nombre infini de valeurs mesurées sont théoriquement possibles, selon le système de mesure. Ex. : densité optique ou pourcentage de valeurs positives dans un essai immuno-enzymatique ou valeurs seuils d'un essai RT-PCR inférieures au nombre maximum de cycles effectués pour l'essai.

Par souci de concision, cette annexe prend en compte les méthodes de validation des tests candidats courantes sans étudier toutes les méthodes statistiques susceptibles d'être utilisées en pratique. Ces méthodes sont décrites pour estimer la précision d'un essai répété à plusieurs reprises (répétabilité et reproductibilité), ses caractéristiques analytiques (sensibilité et spécificité analytiques) et ses caractéristiques diagnostiques (sensibilité diagnostique [SeD], spécificité diagnostique [SpD] et aire sous la courbe d'efficacité du récepteur d'un essai) lorsqu'il est utilisé pour détecter un analyte chez les animaux à titre individuel. Des principes analogues s'appliquent lorsque les tests sont utilisés pour détecter le même analyte dans des groupes d'échantillons constitués naturellement ou artificiellement et provenant d'ensembles d'animaux (troupeaux ou cheptels). Dans ce cas, l'unité épidémiologique est l'ensemble plutôt que l'animal pris individuellement.

Les méthodes statistiques diffèrent si un ou plusieurs tests sont évalués, selon leur échelle de mesure (binaire, ordinale ou continue), si des échantillons indépendants ou dépendants (appariés) sont utilisés et s'il existe un étalon de référence parfaitement précis (référence absolue/« gold standard ») pour comparaison (Wilks, 2001). Des organigrammes destinés à guider le choix de la méthode statistique pour l'évaluation de la précision diagnostique des mesures, telle la sensibilité ou la spécificité, sont représentés aux Figures 1 et 2.

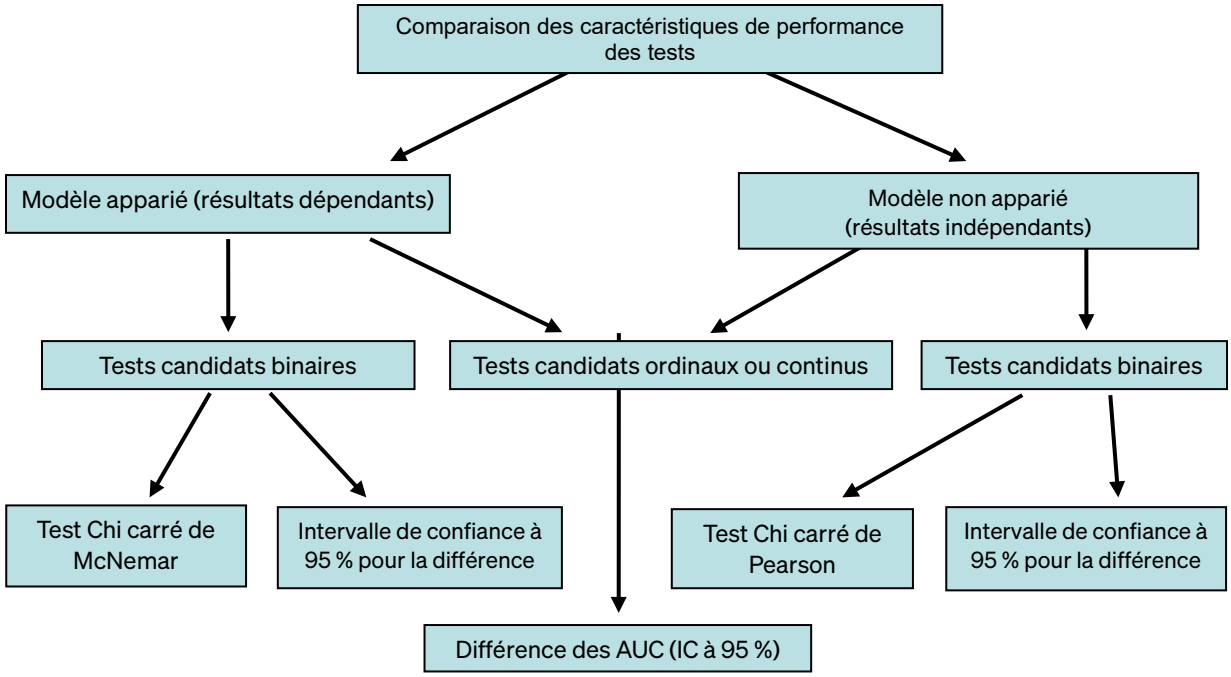
*L'adéquation de l'analyse statistique au modèle de l'étude ne se reflète pas toujours dans la qualité de sa description dans les publications scientifiques. Aussi, les concepteurs et les évaluateurs des tests sont encouragés à suivre la liste des STARD (**Sta**ndards for **R**eporting of **D**iagnostic Accuracy) (Bossuyt* et al., *2003) pour garantir la restitution exhaustive de toutes les informations importantes des études de validation des maladies infectieuses des animaux.*

Pour des instructions sur les données d'analyse de l'incertitude des mesures et pour les données des études de comparaison des méthodes, se référer aux Chapitres 2.2.4 et 2.2.8 respectivement.



Abréviations : SeD = sensibilité diagnostique ; SpD = spécificité diagnostique ; ROC = caractéristiques récepteur-opérateur ; IC = intervalle de confiance ; IP = intervalle de probabilité.

Fig. 1. Organigramme des méthodes d'analyse statistique suggérées lorsqu'un test candidat est évalué individuellement avec ou sans étalon de référence parfait.



Abréviation : AUC = aire sous la courbe ROC (caractéristiques récepteur-opérateur)

Fig. 2. Organigramme pour les méthodes d'analyse statistique suggérées lorsque la sensibilité diagnostique (SeD), la spécificité diagnostique (SpD) et l'AUC de plusieurs tests sont évaluées par rapport à un étalon de référence parfait. Les données ordinales et continues doivent être analysées dans leur format d'origine et les résultats binaires aux seuils recommandés. Les analyses concernant la SeD et la SpD

doivent être effectuées lorsque celles-ci sont disponibles.

A. REPETABILITE D'UN ESSAI DANS UN LABORATOIRE UNIQUE

Une évaluation de la répétabilité intralaboratoire d'un essai (souvent appelée *précision* lorsque la mesure se fait sur une échelle continue) requiert qu'un minimum de trois échantillons dont les concentrations d'analyte se situent à l'intérieur de la plage de fonctionnement de l'essai soient testés à plusieurs reprises par un seul opérateur utilisant un seul lot de kit de test. En principe, ces cycles sont effectués le même jour, mais peuvent aussi s'étaler sur plusieurs jours. L'utilisation de trois ou quatre copies d'un échantillon plutôt que deux est vivement recommandée, car cela exprime mieux la variabilité inhérente des résultats intra-cycle. L'utilisation de plus de deux copies peut ne pas être faisable pour tous les types d'essais, et ce, pour des raisons de coûts (p. ex. détection de l'acide nucléique). Comme décrit dans la Norme de validation de l'OMSA, la variation inter-cycles peut être évaluée sur plusieurs cycles et impliquer deux opérateurs ou plus, à des jours différents. Les deux sections qui suivent décrivent les méthodes destinées à analyser les données continues et binaires pour la répétabilité des essais.

1. Résultats continus

Pour les résultats continus, la méthode la plus simple consiste à estimer l'écart type des copies d'un ensemble d'échantillons représentatif de la plage de fonctionnement de l'essai. Ces résultats doivent être initialement évalués sur un diagramme de dispersion ou un diagramme où la moyenne des copies est représentée par rapport à l'écart type. Pour les essais où l'écart type est proportionnel à la moyenne, le coefficient de variation (CV) intra-échantillon est souvent calculé. Le CV est fréquemment utilisé, même en l'absence de proportionnalité. Dans ce cas, il doit être reporté pour les différents niveaux de concentration de l'analyte cible (faible, modérée ou élevée). Cela est nécessaire, car le CV observé est fréquemment plus élevé lorsque la concentration de l'analyte cible est basse. D'une manière générale, une estimation de l'incertitude des coefficients de variation (intervalle de confiance [IC] à 95 %) doit également être calculée. Lorsque les coefficients de variation sont relativement constants sur la plage des valeurs du test, cela peut se faire en utilisant les résultats de tous les échantillons. Lorsque le coefficient de variation diffère en fonction de la concentration d'analyte, un intervalle de confiance à 95 % doit être calculé séparément pour chaque catégorie d'analyte, sur la base du nombre d'échantillons testés à chaque niveau. Les méthodes de calcul de l'intervalle de confiance pour le coefficient de variation et de la différence de deux coefficients pour des données normales sont décrites dans Donner et Zou (2012).

$$CV = \frac{\text{ET des répliques}}{\text{Moyenne des répliques}}$$

où : CV = Coefficient de variation
ET = Écart type

Si la conception de l'expérience inclut l'évaluation de plusieurs facteurs, comme les différents opérateurs ou les différents jours d'analyse, d'autres méthodes, par exemple des modèles de variance groupée (modèles mixtes), peuvent s'avérer nécessaires, l'objectif étant alors de décomposer la variation en une somme de différentes composantes faciles à interpréter. Les modèles de variance groupée peuvent également être utilisés pour les données de reproductibilité (voir Section B).

2. Résultats binaires

En principe, des résultats quantitatifs doivent être utilisés pour l'évaluation de la précision d'un essai lorsque les données sont disponibles sous cette forme, même si les résultats peuvent être dichotomisés aux fins de l'établissement de rapports. Pour les tests binaires intrinsèques qui fournissent des résultats soit positifs, soit négatifs, les statistiques kappa peuvent être utilisées pour mesurer la concordance des résultats de test par rapport au hasard. Le kappa va de 0 (concordance aléatoire) à 1 (concordance parfaite) mais beaucoup de conjectures sont faites sur la manière d'interpréter les valeurs kappa (Fleiss *et al*., 2003 ; Landis et Koch 1977). Une meilleure concordance est typiquement attendue lorsque les résultats de test sont très éloignés des seuils ; c'est pourquoi certains échantillons aux valeurs intermédiaires/suspectes doivent être testés pour éviter des évaluations trop optimistes de la concordance. Une version pondérée de kappa pour les résultats ordinaux (négatifs, suspects ou positifs) peut être utilisée pour montrer qu'une grande différence (différences de deux catégories) est plus importante qu'une plus petite (p. ex. différence d'une catégorie). Un intervalle de confiance à 95 % doit être rapporté pour les estimations pondérées ou non de kappa (Fleiss *et al.,* 2003).

Tableau 1. Exemples de calculs de kappa pour les résultats binaires

Exemple 1. Calcul de kappa sur la base de résultats de tests répétés classés comme positifs ou négatifs

Résultat du test	*Positif*	*Négatif*
Positif	90	5
Négatif	10	95
	100	100

Kappa = 0,85 (IC à 95 % = 0,78 à 0,92)

Exemple 2. Calcul de kappa sur la base des résultats de tests répétés classés en trois catégories (positifs, suspects ou négatifs)

Résultat du test	*Positif*	*Suspect*	*Négatif*
Positif	80	10	10
Suspect	15	75	10
Négatif	5	15	80
	100	100	100

Kappa = 0,68 (IC à 95 % = 0,61 à 0,75). Kappa pondéré = 0,70 (IC à 95 % = 0,61 à 0,79).

B. REPRODUCTIBILITE DE L'ESSAI ENTRE LABORATOIRES

La précision de l'essai varie selon la routine de sa mise en œuvre, c'est-à-dire selon les différences entre les opérateurs, sites de test, lots de kits ou selon les jours. Le plus souvent, le terme *reproductibilité* s'applique à l'évaluation de la précision de l'essai concerné dans plusieurs laboratoires. Les facteurs maintenus constants devraient être décrits pour permettre l'interprétation des résultats dans le contexte de la situation de test instantanée. Les études de reproductibilité peuvent être faites de manière indépendante des études de répétabilité ou en association avec celles-ci mais elles doivent se faire en aveugle. Comme suggéré dans la Norme de validation de l'OMSA, trois laboratoires au moins doivent tester un minimum de 20 échantillons avec des aliquotes identiques allant à chaque laboratoire.

Le coefficient de corrélation intra-classe représente la similarité ou la corrélation de deux mesures quelconques faites sur le même échantillon. Le CCI prend des valeurs situées entre 0 et 1, les valeurs proches de 1 indiquant une erreur de mesure minimale et, inversement, les valeurs proches de 0 indiquant une quantité importante d'erreurs de mesure.

Les méthodes statistiques pour l'analyse des études de reproductibilité d'un essai interlaboratoires sont similaires à celles utilisées pour l'évaluation de la répétabilité intralaboratoire. Cependant, il peut être jugé important d'évaluer et d'ordonner (en classes, selon le terme consacré) la variabilité des résultats des tests de provenances diverses, cela faisant partie des études interlaboratoires. Si, par exemple, une étude a été conçue pour tester un essai dans trois laboratoires utilisant chacun deux techniciens hautement qualifiés et analysant les échantillons à double avec deux lots de kits, chaque échantillon sera testé 24 fois. Les facteurs sélectionnés (laboratoire, technicien, lot de kits, résultat des copies) peuvent être considérés comme fixes ou aléatoires selon la manière dont ils ont été sélectionnés et de leur degré de représentativité de la population cible. Pour la conception de l'étude, les composantes de variance peuvent être estimées pour chaque classe (exemple : Dargatz *et al.,* 2004) et le coefficient de corrélation intra-classe (intra-cluster, CCI) peut être estimé comme étant une mesure de la similarité des résultats des échantillons (Bartlett et Frost, 2008).

1. Lorsqu'une modification technique est apportée à la méthode de test

Une fois qu'un essai a été validé pour être utilisé dans l'environnement contrôlé d'un laboratoire, son utilisation peut être envisagée dans un environnement complètement différent (par exemple sur le terrain). En raison de l'ampleur des changements, telles les fortes fluctuations de température, fréquentes sur le terrain, il est légitime de prévoir que les deux tests se comporteront différemment selon leur environnement. De fait, on peut s'attendre, dans une étude de ce genre, à ce que les valeurs soient interprétées plutôt comme une erreur de mesure systématique des mesures (ce qui serait le cas si ces valeurs amenaient une sur- ou une sous-estimation des valeurs véritables) que comme une erreur aléatoire (s'appliquant à l'évaluation de l'erreur de mesure

intralaboratoire ou interlaboratoires). Dans l'exemple d'analyses d'échantillons répartis entre le terrain et le laboratoire, la moyenne des différences entre la valeur sur le terrain et la valeur en laboratoire (vraie valeur) pour le même échantillon doit être rapportée avec un IC à 95 %. Si l'intervalle de confiance à 95 % exclut zéro, c'est la preuve d'une déviation systématique des résultats du test, les résultats sur le terrain n'étant pas comparables avec ceux de l'essai validé en laboratoire. Pour valider l'essai sur le terrain, celui-ci devra faire l'objet soit d'une « modification technique », qui sera évaluée dans une étude de comparaison des méthodes (voir Chapitre 2.2.8), soit d'une revalidation complète de son utilisation sur le terrain.

Des approches analogues peuvent être utilisées pour évaluer les changements apportés à une méthode en laboratoire et déterminer s'il en résulte une variation systématique ou aléatoire des résultats.

Exemple : les données qui suivent n'ont pas été publiées, mais ont été obtenues en comparant les valeurs seuils (Ct) de deux méthodes d'extraction (échantillons répartis entre ancienne et nouvelle méthode) pour une réaction en chaîne par polymérase quantitative en temps réel (qPCR) de la langue bleue. Les données (n = 10) représentent les moyennes d'échantillons en double.

Ancienne méthode : 25,6 ; 24,5 ; 21,3 ; 26,8, 25,2 ; 30,2 ; 31,2 ; 32,8 ; 31,8 ; 34,9.

Nouvelle méthode : 23,1 ; 21,0 ; 18,2 ; 25,2 ; 24,7 ; 28,6 ; 30,4 ; 32,2 ; 31,3 ; 34,7.

La différence moyenne entre les deux méthodes (ancienne moins nouvelle) était de –1,49 (IC à 95 % = –2,33 à –0,64) avec une probabilité bilatérale de p = 0.003. Comme l'intervalle de confiance à 95 % exclut zéro, cela indique une valeur Ct systématiquement inférieure lorsque la nouvelle méthode d'extraction est utilisée. Un graphique de Bland–Altman (Bland et Altman, 1999 ; Fig. 3) peut être utilisé pour représenter graphiquement l'évolution de la différence en fonction de la valeur moyenne de l'ancienne et de la nouvelle méthode. Pour ces données, la différence semble diminuer pour les valeurs élevées de Ct, mais l'échantillonnage est de petite taille.

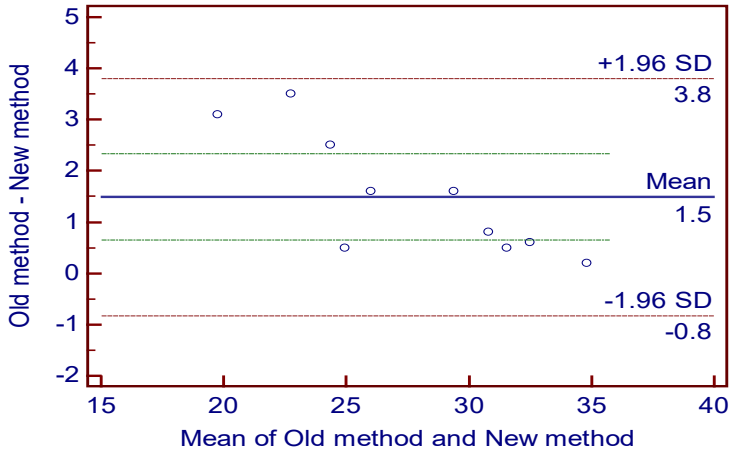


***Fig. 3. Diagramme de Bland-Altman de la différence moyenne* (axe *y*) *des valeurs Ct en fonction de la valeur moyenne de l'ancienne et de la nouvelle méthode* (n *= 10*)**

C. SENSIBILITE ANALYTIQUE (SeA, SYNONYME = LIMITE DE DETECTION : LD)

La sensibilité analytique peut être estimée avec une expérience de dilution jusqu'à extinction, au cours de laquelle des séries de dilution avec une teneur quantifiée d'analyte cible sont intégrées à la matrice d'échantillon appropriée. Cette teneur connue peut être celle d'un étalon de référence interne ou national/international ou d'un échantillon prélevé sur le terrain dont la concentration en analyte a été déterminée. Une analyse parallèle de l'étalon de comparaison peut être faite, mais elle n'est pas indispensable, à moins que l'étude ne vise à comparer une modification mineure apportée à un essai validé par rapport à l'essai original. Cette méthode de dilution jusqu'à extinction peut être utilisée que l'analyte soit mesuré qualitativement ou quantitativement. Dans ce dernier cas, le résultat du test sera reclassé comme positif ou négatif.

La méthode pour analyser les données de limite de détection dépend du modèle expérimental. Prenons par exemple une étude où l'on aura ajouté 10^8 unités formant des colonies (UFC) d'une bactérie à 10 g de fèces pour

obtenir une concentration de 10^7 UFC/g, puis dilué l'échantillon en séries de dilutions décimales jusqu'à 10^1 UFC/g. L'expérience est répétée trois fois. Si toutes les copies à 10^3 UFC/g sont détectées mais aucune à 10^2 UFC/g, la limite de détection pourra être traditionnellement estimée comme étant de 10^3 UFC/g. Si une estimation précise de la LD était nécessaire, une deuxième phase d'expérience pourrait être conçue pour déterminer la LD avec une plus grande certitude en utilisant une série de dilutions plus fines, par exemple au demi, couvrant l'intervalle entre 100 % et 0 % de détection de la première expérience. Le critère de limite de détection est souvent défini comme 95 % ; dans une expérience avec 20 copies, cela correspond à la dilution où 19 répliques sont positives pour l'analyte. L'élément clé est que le critère de probabilité choisi pour la limite de détection (qu'il s'agisse de 95 %, de 50 % ou d'une autre valeur) doit être spécifié et utilisé de manière systématique s'il s'agit de comparer les résultats de plusieurs tests. La limite de détection peut être estimée selon la méthode non paramétrique de Spearman-Kärber ou avec une analyse de régression logistique ou de régression des probits. Plus le nombre de copies est grand pour chaque dilution, plus l'estimation de la limite de détection sera précise.

Exemple : Guthrie *et al.* (2013) a fait une série de dilutions au demi d'un sang de cheval positif au virus de la peste équine (dilution : 10^{-3}) couvrant la plage non linéaire de l'essai. L'extraction a été répétée 25 fois et les échantillons testés par RT qPCR. Les résultats de la qPCR pour les 15 points de dilution ont été utilisés dans une analyse des probits pour calculer la limite de détection à 95 %, soit la concentration donnant un résultat positif à la RT qPCR pour 95 % des copies (Burns et Valdivia, 2008). La limite de détection à 95 % a été estimée comme étant la dilution de 3.02×10^{-6}, ainsi que le montre la Figure 4, et correspond à une quantification cycle de 35,71 pour la qPCR. L'intervalle de confiance de l'estimation n'est pas indiqué.

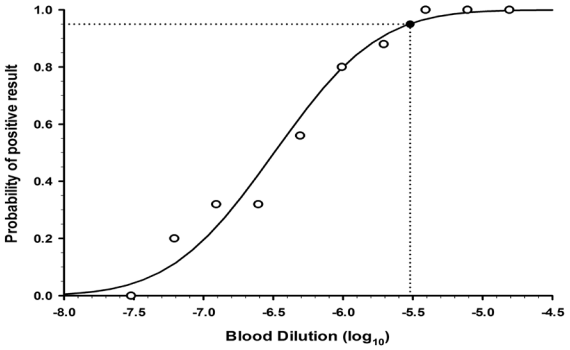


Fig. 4. Estimation de la limite de détection à 95 % du virus de la peste équine dans le sang d'un cheval (*log_{10}*), montrée par la ligne en pointillés

D. SPECIFICITE ANALYTIQUE (SpA)

La spécificité analytique peut être décrite de trois manières au moins : la sélectivité, l'exclusivité (synonyme : profil de réactivité croisée) et l'inclusivité (comme décrit dans la Norme de validation de l'OMSA). Ces deux dernières mesures doivent être mises en regard de la lignée, de l'isolat, de l'espèce ou du genre, selon ce qui convient pour l'analyte cible et l'objectif prévu du test. Pour les tests de dépistage, une spécificité plus large et plus inclusive que pour les tests de confirmation est requise, qui soit à même de différencier des isolats dont la pathogénicité varie, par exemple. Comme le choix des organismes apparentés est subjectif et dépend souvent du type et du nombre d'échantillons, le résultat de l'exclusivité doit être donné qualitativement, c'est-à-dire sous forme de pourcentage d'agents apparentés montrant une réactivité croisée dans l'essai par rapport à une liste d'agents examinés susceptibles de réactivité croisée. De manière analogue, l'inclusivité sera rapportée en pourcentage de sérovars, de souches, de genres ou d'espèces que l'essai a permis de détecter, selon ce qui convient pour l'analyte cible.

E. PERFORMANCE DIAGNOSTIQUE DE L'ESSAI

La performance diagnostique d'un essai se mesure le plus souvent à sa sensibilité (SeD), à sa spécificité (SpD) ou sous forme de mesure combinée de la SeD et de la SpD, par exemple le rapport de vraisemblance des résultats positifs et négatifs. Les rapports de vraisemblance pour les intervalles des résultats de test peuvent aussi se

calculer lorsqu'il est important de conserver les informations sur l'intensité des résultats de test plutôt que de les utiliser sous une forme dichotomisée. Pour plus d'informations sur l'utilisation et le calcul des rapports de vraisemblance, voir Gardner et Greiner (2006) et Gardner *et al.* (2010). Ce dernier article comprend un exemple de toxoplasmose et de calcul de l'intervalle de confiance selon deux méthodes.

L'incertitude statistique concernant les paramètres de performance diagnostique, à savoir SeD et SpD, doit être présentée sous forme d'intervalles de confiance (IC). Traditionnellement, un IC à 95 % est utilisé et sa largeur (précision de la valeur estimée) dépend fortement de la taille de l'échantillonnage utilisé pour l'estimation des paramètres. Il est préférable d'avoir des IC exacts plutôt que des approximations normales, ceux-ci permettant d'éviter les limites supérieures dépassant 100 %.

La SeD et la SpD peuvent être estimées lorsque la méthode de référence ou de comparaison est parfaitement sensible et spécifique ou lorsque l'étalon de référence est imparfait. En général, la plupart des étalons de référence *ante mortem* couramment utilisés dans les laboratoires de diagnostic sont imparfaits ; une nécropsie accompagnée de l'analyse de plusieurs tissus selon des tests secondaires tels que culture et/ou histopathologie est donc souvent nécessaire pour pouvoir considérer les résultats de l'étalon de référence comme justes. Pour la plupart des études de validation concernant les maladies animales, cette dernière option n'est pas réalisable ou ne se justifie pas financièrement, sauf pour un nombre limité d'échantillons.

1. La SeD et la SpD avec un étalon de référence parfait

Le test candidat peut fournir des résultats sur une échelle binaire (dichotomique), ordinale (p. ex. titre) ou continue. Pour ces deux dernières, les résultats doivent être dichotomisés avant de pouvoir calculer la SeD et la SpD ; un seuil doit donc être établi. Un intervalle de confiance à 95 % exactement binomial est recommandé pour la SeD et la SpD (Greiner et Gardner, 2000), car une approximation normale risque de ne pas fournir un IC approprié lorsque les estimations paramétriques sont proches de 1.

Exemple. Essai immuno-enzymatique indirect (I-ELISA)

		Nombre d'animaux	
		Séropositifs connus (369)	Séronégatifs connus (198)
Résultats du test	Positif	287 VP	1 VP
	Négatif	82 FN	197 VN
		Sensibilité diagnostique* VP/(VP + FN) **77,8 %** (73,2 – 81,9 %)*	Spécificité diagnostique* VN/(VN + FP) **99,5 %** (97,2 – 99,9 %)*

VP et FP = vrais positifs et faux positifs, respectivement

VN et FN = vrais négatifs et faux négatifs, respectivement

*Limites de confiance binomiales exactes à 95 % pour la SeE et la SpD

Lorsque l'étalon de référence n'est pas utilisé pour tous les résultats de test positifs et négatifs (vérification partielle), une correction des estimations de la SeD et de la SpD doit être faite de la manière décrite par Greiner et Gardner (2000) afin de prendre en compte les différentes probabilités d'échantillonnage dans les groupes séropositifs et séronégatifs.

Pour les essais qui fournissent des résultats ordinaux (p. ex. valeurs de titrage) ou continus (p. ex. rapport des valeurs des échantillons de test sur celles des échantillons témoins positifs dans un ELISA), les estimations de la SeD et de la SpD doivent être complétées par des estimations de l'aire sous la courbe des caractéristiques récepteur-opérateur (ROC). L'analyse ROC offre une approche indépendante des seuils pour l'évaluation de la précision globale d'un test lorsque les résultats sont mesurés en valeurs ordinales ou continues. L'aire sous la courbe ROC fournit une estimation numérique unique de la précision globale, allant de 0,5 (test sans utilité) à 1 (test parfait). La principale justification de l'analyse ROC est que les valeurs seuils pour l'interprétation du test

peuvent changer selon l'objectif de l'analyse (p. ex. dépistage versus confirmation), la prévalence de l'infection, le coût des erreurs d'analyse et la disponibilité d'autres tests. Des descriptions détaillées de l'analyse ROC sont présentées ailleurs (Gardner et Greiner, 2006 ; Greiner *et al.,* 2000 ; Zweig et Campbell, 1993). Lorsque plusieurs tests ordinaux ou continus sont comparés, la différence de leurs aires sous la courbe avec un intervalle de confiance à 95 % doit être calculée. Les méthodes pour calculer ces différences ne sont pas les mêmes pour les échantillons indépendants ou dépendants ; elles sont appliquées dans de nombreux programmes statistiques (Gardner et Greiner, 2006). Les Figures 5 et 6 montrent des exemples de diagrammes de dispersion des résultats d'un ELISA unique et de courbes ROC pour deux ELISA.

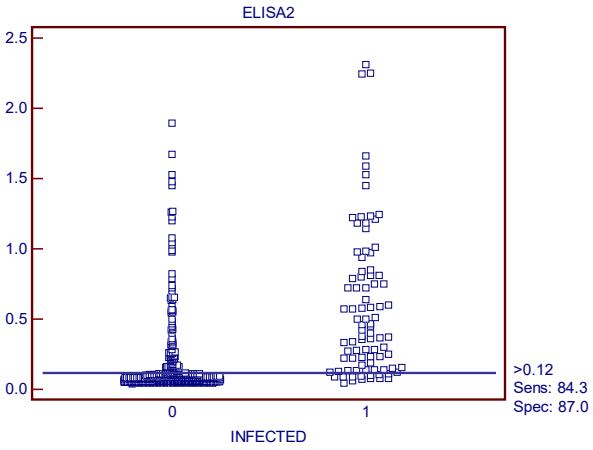


Fig. 5. Diagramme de dispersion des résultats d'un ELISA pour les animaux non infectés (Code = 0) ou infectés (Code = 1).

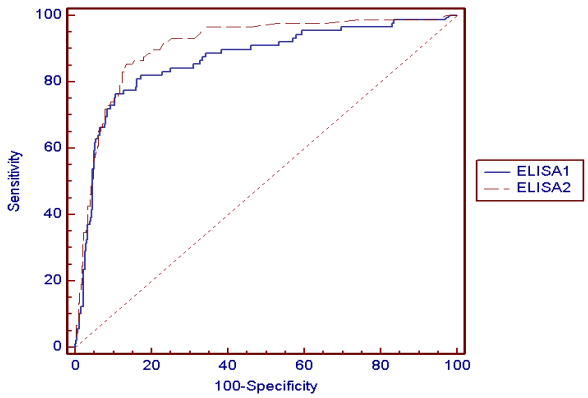


Fig. 6. Courbe ROC pour deux ELISA.

En l'absence d'étalon de référence parfait, il est également possible d'estimer l'AUC au moyen de modèles de structure latente. Des modèles de structure latente peuvent par exemple être appliqués à des données normalement distribuées provenant de deux tests dépendants (voir Choi *et al.,* 2003) et en utilisant des approches semi-paramétriques (Branscum *et al.,* 2008). En raison de leur complexité, les modèles de structure latente pour les données continues incluant les données censurées ou tronquées, ce qui arrive avec les essais de PCR en temps réel, ne sont pas décrits dans ces lignes directrices pour la validation. Toutefois, des modèles de structure latente pour les résultats de test binaires ainsi qu'un exemple de leur utilisation sont décrits à la Section E.3.

2. Comparaison des estimations de la SeD et de la SpD avec un étalon de référence parfait pour deux tests

Souvent, les chercheurs aimeraient pouvoir comparer les valeurs de SeD dans des sous-populations d'animaux infectés, p. ex. des animaux infectés cliniquement par rapport aux animaux infectés subcliniquement, ou les valeurs de SpD dans des zones géographiques différentes. Comme il s'agit d'échantillons indépendants, les comparaisons peuvent se faire de manière statistique avec le test du Chi carré de Pearson pour l'homogénéité.

Sinon, il est possible de calculer un intervalle de confiance à 95 % séparé et un intervalle de confiance à 95 % pour une différence de deux proportions. Lorsque la SeD (ou la SpD) de deux tests est comparée sur le même ensemble d'échantillons infectés (ou non) dans un modèle apparié, les résultats de test ne sont plus indépendants. Des méthodes statistiques comme le Chi carré de McNemar peuvent être utilisées pour tester l'hypothèse de sensibilités (ou de spécificités) égales si le test est effectué sur les mêmes échantillons.

Exemple : cinq titrages d'anticorps ont été évalués pour le diagnostic de la paratuberculose bovine chez les vaches laitières de troupeaux connus pour être infectés ou non infectés, selon les résultats de cultures fécales et selon les antécédents du troupeau. Les tableaux de données suivants ont été générés sur la base des données originales avant d'être publiés dans Collins *et al*. (2005). Dans la publication, un troupeau a été exclu de l'analyse. Cet exemple est utilisé à des fins de démonstration et montre la disposition des tableaux pour le calcul de la SeD, de la SpD et pour leur évaluation statistique.

Infectés

	T_{2+}	T_{2-}	
T_{1+}	124	74	198
T_{1-}	8	243	251
	132	317	449

Sensibilité de T1 = 198/449 = 44,1 %

Sensibilité de T2 =132/449 = 29,4 %

Non infectés

	T_{2+}	T_{2-}	
T_{1+}	3	27	30
T_{1-}	16	366	382
	19	393	412

Spécificité de T1 = 382/412 = 92,7 %

Spécificité de T2 = 393/412 = 95,4 %

Sur la base de deux tests personnalisés de Chi carré selon McNemar, les sensibilités diffèrent significativement ($p < 0.0001$) mais pas les spécificités ($p = 0.126$). Les covariances de la sensibilité et de la spécificité (voir Gardner *et al*., 2000 pour les détails) peuvent aussi être calculées pour déterminer la probabilité conditionnelle que les tests soient indépendants ou dépendants, selon le statut infectieux. Pour ces données, la covariance de la sensibilité (calculée en utilisant le tableau de gauche : « Infectés ») est de 0,147 ($p < 0.0001$ selon le Chi carré de Pearson), montrant une forte dépendance des deux tests lorsqu'ils sont utilisés sur des animaux infectés. La covariance de la spécificité (calculée en utilisant le tableau de droite : « Non infectés ») est de 0,004 ($p = 0,152$ selon le Chi carré de Pearson), montrant qu'il n'y pas de dépendance significative.

Un exemple supplémentaire basé sur des données de toxoplasmose porcine est présenté dans Gardner *et al*. (2010).

3. DSe et DSp sans étalon de référence parfait

Les progrès des méthodes statistiques, notamment le développement des modèles de structure latente (« no gold standard »), permettent désormais aux chercheurs de se libérer des hypothèses restrictives d'un test de référence parfait et d'estimer la précision du ou des tests candidats ainsi que des étalons de référence avec les mêmes données (Enoe *et al.,* 2000 ; Hui & Walter, 1980).

Les modèles de structure latente, qu'ils utilisent la probabilité maximale ou les méthodes bayésiennes, peuvent être utilisés pour estimer la SeD et la SpD lorsque des résultats de tests conjoints sont disponibles, provenant de plusieurs tests appliqués à des animaux de plusieurs populations (troupeaux ou zones géographiques). Tous les modèles de structure latente ne permettent pas la déduction statistique des estimations de la SeD et de la SpD. Un modèle est utilisable s'il permet théoriquement de déterminer la vraie valeur des paramètres du modèle après avoir fourni un nombre infini d'observations. En substance, cela revient à avoir un seul ensemble de valeurs pour les paramètres faisant l'objet des tests (SeD, SpD). Les méthodes bayésiennes sont particulièrement adaptées aux situations où l'on dispose d'informations préalables à propos de la SeD et/ou de la SpD ainsi que lorsqu'il n'y a pas de problèmes d'estimation manifestes (Branscum *et al*., 2005).

Le modèle le plus simple de structure latente appliqué à une population est celui de trois analyses simultanées, conditionnellement indépendantes, effectuées sur les mêmes échantillons. La contrainte d'indépendance des trois tests peut être difficile à remplir en pratique à moins que l'analyte cible ne diffère entre les tests. L'approche couramment utilisée en santé animale est donc d'effectuer deux tests sur tous les échantillons d'animaux de deux

populations car c'est moins coûteux et les hypothèses d'indépendance conditionnelle sont potentiellement plus raisonnables. Le modèle de deux tests dans deux populations implique également l'hypothèse d'une sensibilité et d'une spécificité constante parmi les deux populations ainsi que des prévalences distinctes. L'hypothèse d'une sensibilité constante peut être difficile à vérifier et est peu susceptible d'être correcte si une population présente des animaux malades cliniquement et que l'autre ne présente que des cas subcliniques, plusieurs études publiées ayant montré que la sensibilité du test est plus grande chez les animaux cliniquement touchés. Si l'une des deux populations est connue pour être indemne de l'agent pathogène (prévalence de zéro), tandis que l'autre population est connue pour avoir une prévalence différente de zéro, la première population peut être utilisée pour estimer la SpD, ce qui facilitera l'estimation de la SeD dans la population infectée.

Parmi les maladies de la Liste de l'OMSA pour lesquelles la SeD et la SpD ont été estimées avec des méthodes bayésiennes figurent la brucellose ovine (Praud *et al*., 2012), la fièvre Q (Paul *et al*., 2013), la trypanosomiase (Bronsvoort *et al*., 2010), la tuberculose bovine (Clegg *et al*., 2011), la fièvre aphteuse (Bronsvoort *et al*., 2006), la peste équine (Guthrie *et al*., 2013) et l'infection par le virus de l'anémie des salmonidés (Caraguel *et al*., 2012).

Le logiciel WinBUGS[1] permet la mise en œuvre facile des méthodes de Monte-Carlo par chaîne de Markov pour les estimations bayésiennes (Lunn *et al.,* 2000) et des analyses simples de probabilité maximale peuvent être faites avec une interface disponible en ligne (Poulliot *et al.,* 2002). Les informations primaires sur les paramètres du modèle utilisé dans les analyses bayésiennes peuvent modifier les estimations finales, selon le poids relatif de la preuve apportée par les valeurs primaires (degré d'incertitude des valeurs primaires) et par les données (incertitude attribuable au caractère fini de la taille de l'échantillon). Les sources d'informations primaires doivent dont être bien documentées dans les analyses bayésiennes et il peut être souhaitable de répéter l'analyse en utilisant des valeurs primaires non informatives pour tous les paramètres lorsque le modèle est identifiable.

Probabilité maximale : méthode d'estimation des valeurs les plus probables pour les paramètres intéressants, basée sur la valeur de la fonction de probabilité pour les données.

Méthodes bayésiennes : incorporent des informations ou des connaissances primaires importantes à propos d'un ou de plusieurs tests en plus de la fonction de probabilité pour les données. Pour une taille élevée d'échantillonnage, la probabilité maximale et les méthodes bayésiennes amènent à des déductions similaires.

Il est important de noter que l'analyse des structures latentes ne peut pas corriger les biais inhérents aux études mal conçues. Ces méthodes doivent être utilisées avec précaution et inclure une évaluation complète des hypothèses sous-jacentes (dépendance conditionnelle, sensibilité et spécificité constantes parmi les populations et prévalences distinctes), des effets de l'utilisation des distributions primaires choisies sur les déductions secondaires, comme décrit au paragraphe précédent, et de la convergence des chaînes de Markov dans une analyse bayésienne (Toft *et al.,* 2005).

Exemple : Guthrie *et al*. (2013) ont estimé la SeD et la SpD d'une PCR quantitative en temps réel et d'un isolement viral conventionnel pour la détection du virus de la peste équine dans des échantillons de sang complet à l'aide d'un modèle de structure latente bayésien à deux tests et à deux populations. Deux populations de pur-sang sud-africains (503 cas suspects de peste équine et 503 chevaux en bonne santé provenant de la zone où le virus de la peste équine est maîtrisé) ont été testées par PCR et isolement viral. Pour les 503 cas suspects, les résultats des tests conjoints étaient : PCR+VI+ ($n = 156$). PCR+VI– ($n = 184$), PCR–VI+ ($n = 0$) et PCR–VI– ($n = 163$) Les 503 chevaux en bonne santé étaient tous PCR–VI–. Différents modèles (indépendance et dépendance conditionnelles) ont été ajustés aux données et une seconde population de chevaux en bonne santé a été incluse dans certaines analyses.

Les modèles ont été analysés dans WinBUGS 1.4.3 (Lunn *et al.,* 2000), les 5000 premières itérations ont été jetées et les 50 000 itérations suivantes ont été utilisées pour les déductions secondaires (médianes et intervalle de probabilité à 95 % pour la SeD et la SpD). La convergence des modèles a été évaluée par inspection visuelle des tracés représentant les valeurs itérées et par la réalisation de plusieurs chaînes à partir des valeurs de dispersion initiales. Le modèle d'indépendance conditionnelle personnalisé avec les distributions bêta primaires non informatives (1, 1) de la SeD et de la SpD des deux tests a fourni des résultats quasiment identiques à ceux fournis par le modèle utilisant une distribution bêta primaire hautement informative (9999,1) pour la SpD de la neutralisation virale. Les valeurs moyennes estimées et les intervalles de probabilité à 95 % entre parenthèses (parfois appelés intervalles crédibles) provenant du modèle d'indépendance conditionnelle avec des distributions primaires non informatives se présentent ainsi :

1 Accessible sous : http://www.mrc-bsu.cam.ac.uk/bugs/winbugs/contents.shtml

sensibilité de la PCR = 0,996 (0,977–0,999)

spécificité de la PCR = 0,999 (0,993–1,0)

sensibilité de la neutralisation virale = 0,458 (0,404–0,51)

spécificité de la neutralisation virale = 0,999 (0,998–1,0)

Ces résultats montrent une SeD de la PCR deux fois supérieure à celle de la neutralisation virale et une SpD comparable entre les deux tests. Pour une description complète de la méthode de modélisation, voir Guthrie *et al.* (2013).

4. Comparaison des valeurs estimatives de la SeD et de la SpD pour deux tests sans étalon de référence parfait

Si une méthode bayésienne est utilisée dans WinBUGS pour analyser les données de tests conjoints provenant de plusieurs populations, la différence de sensibilité (ou de spécificité) peut facilement être estimée et la probabilité que la sensibilité (ou la spécificité) de l'un des tests excède celle de l'autre peut être estimée avec la fonction STEP.

Exemple : pour les résultats des données de Guthrie *et al.* (2013) à la Section E.3, les intervalles de probabilité à 95 % de la SeD ne se chevauchent pas, tandis qu'il y a un fort chevauchement des intervalles de probabilité à 95 % de la SpD. Les valeurs de probabilité correspondantes obtenues avec la fonction STEP étaient de 1 et de 0,24 respectivement. Ces valeurs apportent la certitude que les SeD diffèrent, mais la probabilité que les SpD diffèrent est faible (moins de 0,5).

RÉFÉRENCES

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**, 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., GAGNÉ 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**, 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**, 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.

*
* *

N. B. : ADOPTE POUR LA PREMIERE FOIS EN 2014.