CHAPITRE 2.2.8.

COMPARABILITE DES EPREUVES SUITE A DES CHANGEMENTS INTRODUITS DANS UNE METHODE DE TEST VALIDEE

INTRODUCTION

Il existe de nombreuses raisons expliquant qu'un essai validé subisse des changements au cours du temps. Le remplacement de réactifs épuisés constitue probablement l'exemple le plus courant de changement mineur introduit dans un essai validé (*voir Chapitre 1.1.6.* Validation des épreuves diagnostiques pour les maladies infectieuses des animaux terrestres*, Figure 1, Mise au point d'un essai et processus de validation*). *Des changements mineurs peuvent être dus : à la disponibilité de réactifs moins chers ou meilleurs (réactifs d'extraction de l'acide nucléique pour les tests moléculaires), au besoin d'augmenter la normalisation* (*plaques pour ELISA pré-enduites plutôt qu'enduites par l'opérateur*), *aux exigences accrues en matière de débit* (*manipulation robotisée plutôt que manuelle*), *etc. Ce type de changements mineurs requiert généralement une étude expérimentale pour évaluer si les caractéristiques de performance de l'essai validé restent comparables avec la nouvelle procédure* (*Tableau 1, Figure 1*). *D'autres variables extérieures à l'essai peuvent nécessiter une vérification, telle la nature de la population, de l'espèce ou de l'individu cible. Par exemple, les diagnosticiens de laboratoire expérimentés se montreront prudents lorsqu'il s'agira de recourir à un ELISA compétitif pour le titrage d'anticorps contre la brucellose chez des bovins en Amérique latine si l'épreuve a été spécifiquement validée pour le bétail au Canada* (*Gall* et al., *1998*). *Les essais sont souvent appliqués à une espèce autre que celle pour laquelle ils ont initialement été validés, p. ex. poulets domestiques versus oiseaux sauvages ou bovins à viande versus vaches laitières. Parmi les autres changements, citons l'utilisation d'échantillons de test différents, p. ex. écouvillon trachéal versus écouvillon cloacal d'oiseaux pour le diagnostic de l'influenza aviaire au moyen d'épreuves moléculaires. Dans ces circonstances, une étude de vérification sera nécessaire pour valider les caractéristiques de performance du test compte tenu des nouvelles conditions.*

> La **validation** est un processus qui détermine l'adéquation d'un essai correctement développé, optimisé et normalisé, à un objectif prévu.
>
> La **vérification** apporte la preuve que les caractéristiques de performance, c'est-à-dire l'exactitude et la précision de l'essai validé, sont comparables lorsque l'essai est utilisé dans un autre laboratoire.
>
> La **comparabilité** est le terme privilégié pour signaler que les caractéristiques de performance d'un nouveau test ayant subi un changement mineur sont aussi bonnes que celles du test validé, dans des limites statistiques définies.
>
> **L'équivalence** a été historiquement utilisée dans certains laboratoires diagnostiques pour les études de comparabilité. Ce terme implique toutefois des exigences plus rigoureuses que l'adéquation à l'objectif prévu et a également une signification statistique spécifique. C'est pourquoi il n'est pas utilisé dans le présent chapitre.

Une controverse existe quant à ce qui constitue un changement « mineur » ou un changement « majeur » pour une épreuve diagnostique. Certains changements sont considérés comme majeurs parce que la base biologique de l'essai s'en trouve fondamentalement modifiée ; ainsi, des changements évolutifs ou des mutations dans la composition de l'acide nucléique d'un agent pathogène nécessiteront l'ajustement des amorces et des sondes. De manière analogue, le passage d'un format d'ELISA indirect à un format compétitif utilisant un anticorps monoclonal hautement spécifique est considéré comme un changement majeur, justifiant une réévaluation complète de l'essai. Le Tableau 1 fournit quelques exemples de changements mineurs ou majeurs fréquents dans les tests de dépistage des anticorps ou de l'acide nucléique. Des études de comparabilité rigoureuses et bien conçues fournissent une

évaluation objective permettant de déterminer si l'essai reste comparable à l'essai validé et adapté à l'usage prévu après introduction d'un changement mineur. Les résultats de l'étude expérimentale permettront de déterminer si l'essai candidat nécessite une revalidation complète et s'il peut ou non être utilisé en toute confiance.

A. MISE SUR PIED D'EXPERIENCES COMPARATIVES

Tableau 1. Exemples de changements dans les épreuves diagnostiques

Type de changement	Modifications de l'essai	Changements dans la population ou dans l'échantillon cible
Mineur	Remplacement d'un réactif épuisé : témoin positif, nouveau lot d'antigène, plaques, conjugué (ELISA), etc. Changement d'instrument ou de plateforme : lecteur ELISA, incubateur/mélangeur, thermocycleur (PCR), etc. Passage de plaques ELISA enduites individuellement à des plaques pré-enduites Passage d'une manipulation manuelle à une manipulation robotisée (ELISA, NAD) Changement dans la procédure d'extraction de l'acide nucléique (NAD) Utilisation d'une ou de plusieurs amorces ou sondes modifiées (substitution partielle de séquences, p. ex. dégénérescence) Conditions de réaction modifiées pour la PCR avec la ou les mêmes amorces et sondes Ajout d'une sonde supplémentaire dans la zone amplifiée Modification chimique de la sonde (NAD)	
Majeur	Remplacement d'un antigène recombinant par un antigène issu d'une culture cellulaire dans un ELISA Passage d'un ELISA indirect à un ELISA de compétition utilisant un anticorps monoclonal spécifique Changement des amorces et des sondes pour des cibles différentes dans plusieurs zones des mêmes gènes ou de gènes différents (NAD)	Espèces différentes, p. ex. bovins versus buffles, poulets domestiques versus oiseaux sauvages Type d'échantillon différent, p. ex. écouvillon trachéal plutôt que cloacal, sang plutôt que sperme, tissus ou organes divers

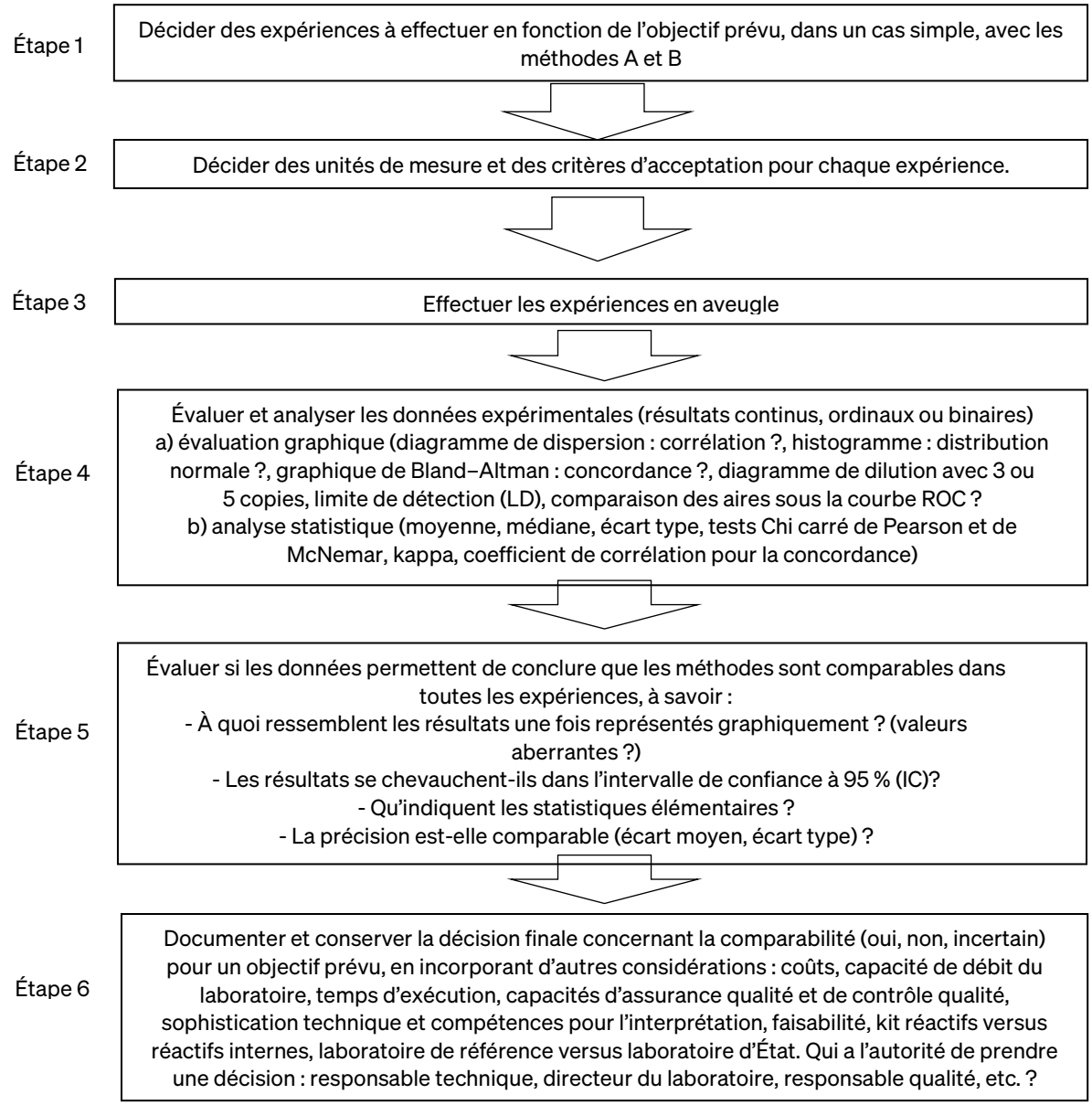
ELISA = dosage immuno-enzymatique ; PCR = réaction de polymérisation en chaîne ;
NAD = test de détection de l'acide nucléique (« nucleic acid detection »).

Lors de la mise sur pied d'une expérience comparative, la procédure doit être axée sur l'objectif de l'essai (Figure 1, étape 1). Par exemple, les tests de dépistage requièrent une sensibilité diagnostique élevée et il est important de comparer leurs limites de détection. Il convient de déterminer une série de dilutions adéquate ainsi que le nombre de copies d'un échantillon témoin. Une quantité suffisante de matériel témoin bien caractérisé doit être produite, répartie en aliquotes et convenablement stockée.

Si l'objectif est d'évaluer et de comparer la répétabilité, il est nécessaire d'analyser des copies bien caractérisées d'échantillons témoins présentant des concentrations en analyte différentes et couvrant la plage de fonctionnement attendue pour l'essai (concentrations élevées, modérées et basses). Pour des raisons pratiques, l'exemple donné dans le présent chapitre utilise uniquement un témoin faiblement positif. Une bonne pratique consiste à inspecter visuellement la corrélation des résultats entre les deux méthodes. Par exemple, un diagramme de dispersion ou un histogramme sont faciles à réaliser et fournissent des informations immédiates sur le type de corrélation et la distribution des données (Figures 2 et 3). Des statistiques élémentaires aident à déterminer les limites supérieures et inférieures et à évaluer les résultats, par exemple pour la limite de détection

ou la répétabilité (Tableaux 2 et 3). Un graphique de Bland-Altman constitue une approche plus sophistiquée pour rapprocher et analyser les résultats d'une étude comparative (Figure 4 et Tableau 4).

Fig. 1. Facteurs à prendre en compte pour les études comparatives de tests diagnostiques.



La Figure 5 est un exemple de la manière de rendre des expériences comparatives plus efficaces en évaluant différents paramètres simultanément sur une seule plaque, à savoir la sensibilité analytique au moyen de dilutions progressives d'un analyte cible en trois copies, puis la spécificité analytique au moyen de quelques échantillons négatifs (non porteurs de l'analyte cible) et, enfin, la sensibilité analytique au moyen d'échantillons provenant d'animaux infectés avec différentes concentrations d'analyte. L'utilisation de copies dans le même cycle et entre les cycles permet d'estimer la répétabilité (Figure 5 et 6).

Il est important de convenir de critères d'acceptation pour évaluer le résultat de l'expérience (Figure 1, étape 2), ce qui signifie que les intervalles de confiance peuvent être asymétriques pour permettre une meilleure performance du nouvel essai. Dans ce chapitre, une estimation conventionnelle de la confiance à 95 % est considérée comme acceptable pour une expérience portant sur la limite de détection (Figures 6 et 7). Pour comparer la répétabilité, il est possible d'utiliser l'écart moyen et l'écart type ou les résultats directs du test (plus ou moins une marge donnée) comme critères d'acceptation (Tableaux 3 et 4 et Figure 4). Les résultats d'un panel d'individus infectés et non infectés fournissent des informations sur la sensibilité et la spécificité diagnostiques comparatives (Figure 8 et Tableaux 5 et 6).

Les données utilisées dans les Figures 2, 3 et 4 et dans les Tableaux 2, 3 et 4 sont produites au moyen des résultats de tests répétés d'un témoin faiblement positif dans deux essais TaqMan ciblant les gènes M (essai M1) et N (essai N1) du virus Hendra.

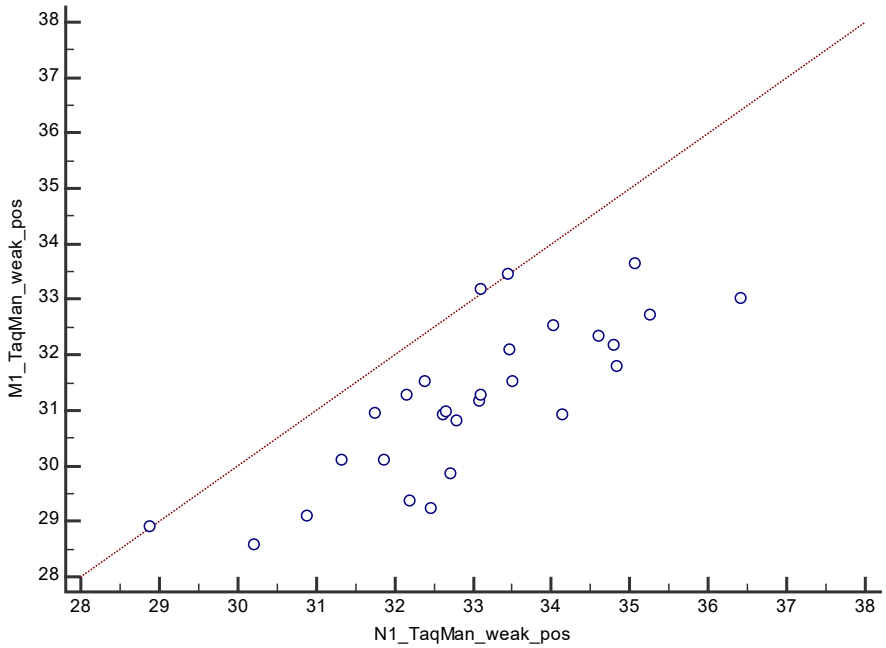
Les données des expériences concernant les limites de détection ainsi que l'aspect des plaques des Figures 5, 6 et 7 sont fictifs. Les données pour les courbes d'efficacité (*receiver operating characteristics*, ROC) de la Figure 8 et des Tableaux 5 et 6 proviennent d'expériences comparatives de différents tests ELISA portant sur l'influenza chez des porcs.

Le présent chapitre donne un aperçu de différentes approches pour concevoir les expériences et interpréter les résultats des études comparatives des essais (sensibilité analytique [limite de détection] et diagnostique ; spécificité analytique et diagnostique ainsi que répétabilité). Il existe de nombreux tests fondamentalement différents, mais les exemples donnés proviennent ou se réfèrent à des expériences de détection d'acide nucléique ou à des dosages immuno-enzymatiques (ELISA). Les principes décrits dans ce chapitre peuvent être considérés comme s'appliquant de manière analogue aux autres tests.

B. INSPECTION VISUELLE

Initialement, un **diagramme de dispersion** est utile pour évaluer visuellement la corrélation entre deux méthodes : le rapport est-il linéaire ou logarithmique ? Voit-on des valeurs aberrantes, des valeurs manquantes, des artefacts ? L'exemple de la Figure 2 utilise les résultats de tests répétés d'un témoin faiblement positif dans deux essais TaqMan ciblant les gènes M (essai M1) et N (essai N1) du virus Hendra. Les résultats montrent que les données pour l'essai N1 sont décalées en bas vers la droite par rapport à l'essai M1, ce qui indique des valeurs systématiquement plus élevées pour l'essai N1 que pour l'essai M1. Le diagramme de dispersion ne fournit cependant pas d'informations sur la concordance des deux tests. Il y a une exception à cette règle : si tous les résultats des deux tests tombent le long de la ligne diagonale à 45 %, la concordance est de 100 %. Sur la Figure 2, seuls trois résultats tombent sur la ligne diagonale. Dans le présent document, la concordance est définie comme un ensemble de valeurs d'un test candidat tombant à l'intérieur de l'intervalle de confiance à 95 % des résultats du test établi, après des analyses répétés du même échantillon témoin bien caractérisé. La corrélation mesure la solidité de la relation entre les mesures de ces tests et elle est exprimée par la valeur p.

Fig. 2. Diagramme de dispersion d'un échantillon témoin faiblement positif testé 28 fois dans deux essais TaqMan Hendra différents, M1 et N1 (résultats exprimés sous la forme de valeurs seuils [Ct]).



Une analyse complémentaire des résultats est faite dans le Tableau 2 ci-dessous, r (coefficient de corrélation) étant égal à 0,8 et indiquant une forte corrélation positive entre les deux méthodes ; $p < 0.0001$ indique que la

probabilité que cette association soit due au hasard est très faible et l'intervalle de confiance à 95 % indique que, lorsque ces méthodes sont utilisées sur un sujet similaire et dans des conditions similaires, nous pouvons avoir confiance à 95 % dans le fait que la véritable valeur inconnue de r se situe entre 0,61 et 0,9.

Tableau 2. Analyse statistique d'un échantillon témoin faiblement positif testé 28 fois dans deux essais TaqMan Hendra, M1 et N1

Variable Y	TaqMan M1faiblement positif
Variable X	TaqMan N1 faiblement positif
Taille de l'échantillonnage	28
Coefficient de corrélation r	0,8015
Niveau de signification	$p < 0{,}0001$
Intervalle de confiance à 95 % pour r	0,6112 à 0,9042

La Figure 3 représente un **histogramme** permettant de détecter une déviation vers la gauche ou vers la droite ou d'autres caractéristiques importantes telle une distribution bimodale.

Fig. 3. Histogramme d'un échantillon témoin faiblement positif testé 28 fois dans deux essais TaqMan Hendra, M1 et N1 (résultats exprimés sous la forme de valeurs seuils [Ct]).

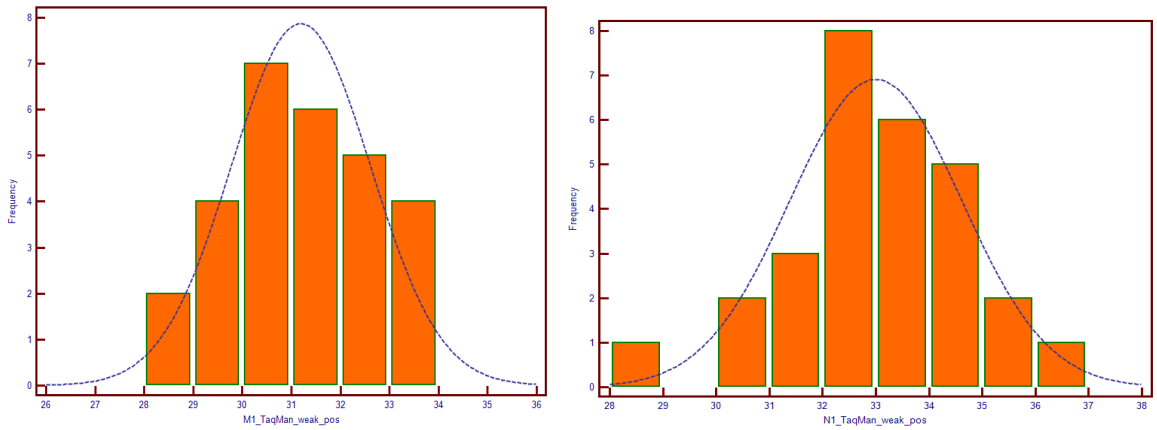


Tableau 3. Analyse statistique d'un échantillon témoin faiblement positif testé 28 fois dans deux essais TaqMan Hendra, M1 et N1 (résultats exprimés sous la forme de valeurs seuils [Ct])

	Taqman M1 faiblement positif	**Taqman N1 faiblement positif**
Taille de l'échantillonnage	28	28
Valeur la plus basse	28.58	28.88
Valeur la plus haute	33.63	36.42
Moyenne	31.19	33.00
Médiane	31.22	32.94
Écart type (ET)	1.42	1.62

C. RÉPÉTABILITÉ

La variabilité d'un essai peut être évaluée au moyen de copies d'un échantillon témoin interne utilisé sur plusieurs cycles dans le temps. Dans cet exemple, la répétabilité a été comparée pour deux essais TaqMan Hendra différents ciblant les gènes N et M et utilisant un témoin interne faiblement positif après 28 cycles effectués par le

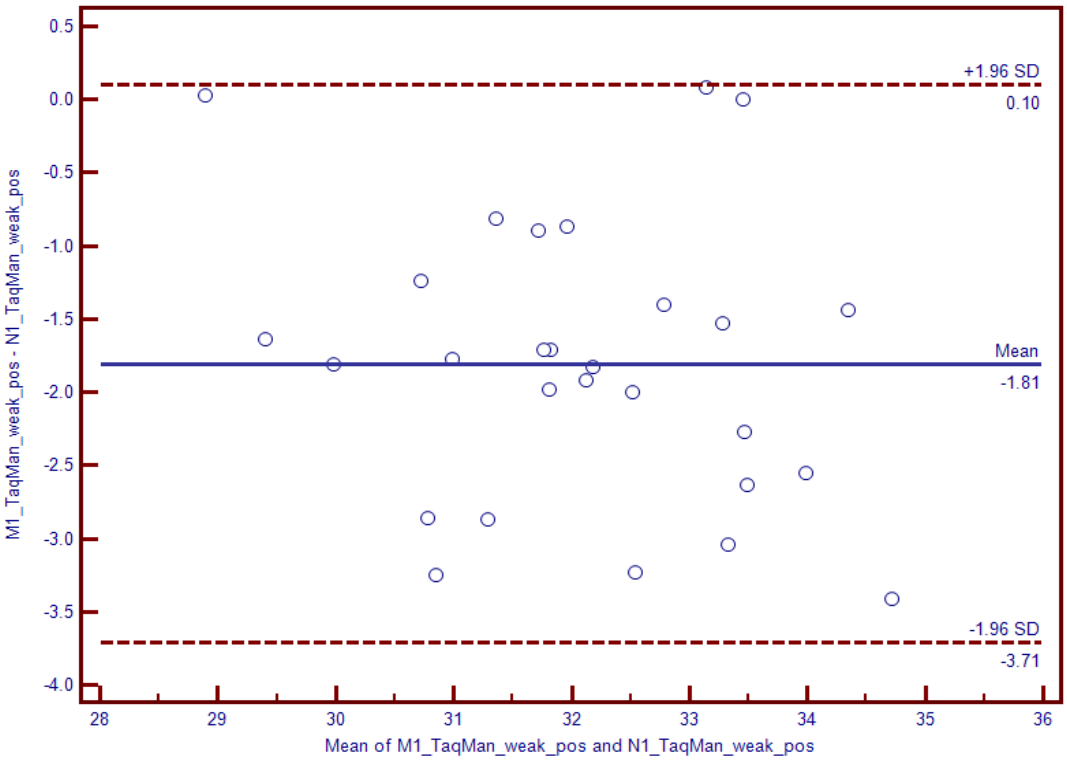
même opérateur pendant 14 jours répartis sur une période de 18 jours. Les résultats seuils (Ct) ont été synthétisés sous forme de moyenne et d'écart type (Tableau 3).

Comme les estimations reposent sur un échantillon témoin unique, aucune comparaison formelle n'est nécessaire. Par contre, si les critères d'acceptation prédéfinis pour les deux essais avaient été, disons, de 2 à 3 ET ou de ±2 ou 3 valeurs Ct, les deux essais auraient alors été considérés comme comparables.

D. GRAPHIQUE DE BLAND–ALTMAN

Un graphique des différences moyennes de Tukey constitue une manière très efficace de montrer et d'analyser simultanément les résultats d'une étude comparative où deux mesures différentes sont faites sur chaque échantillon (modèle dit de paires appariées) (Bland et Altman, 1999, 2007 ; Kozk & Wnuk, 2014). Ce graphique est utile pour mettre en évidence le rapport entre les différences et les moyennes, pour évaluer chaque biais systématique et pour identifier d'éventuelles aberrations. Les valeurs Ct moyennes obtenues avec les méthodes A (essai TaqMan M1) et B (essai TaqMan N1) sur toute la plage des résultats sont reportées sur l'axe des x et les différences des valeurs moyennes, soit A moins B, sur l'axe des y (Figure 4, Tableau 4). Dans notre exemple, le TaqMan M1 est comparé avec le TaqMan N1 en utilisant les résultats d'un témoin faiblement positif après 28 cycles. Toutes les différences sont négatives du fait que les valeurs moyennes pour le TaqMan N1 (Ct 33) sont supérieures à celles du TaqMan M1 (31,99) (Tableau 3). En soustrayant N1 de M1, on constate que presque toutes les valeurs sont négatives, ce qui indique un biais systématique, la différence moyenne entre N1 et M1 étant de –1,81 (biais). Les lignes horizontales sont tracées au niveau de la différence moyenne (–1,81) et aux limites de concordance, définies comme la différence moyenne plus ou moins 1,96 fois l'écart type des différences qui, exprimé en différences Ct moyennes, va de 0,1 à –3,71. Les résultats du Tableau 4 montrent que l'intervalle de confiance à 95 % pour la différence moyenne (–2,18 à –1,43) exclut zéro et qu'il n'est donc pas possible de conclure que les deux méthodes sont comparables. L'ajustement des seuils peut aider à contrebalancer ce biais systématique apparent entre les deux méthodes, une fois ces résultats corroborés avec un grand nombre d'échantillons sur toute la plage de résultats attendue.

Fig. 4. Graphique de Bland–Altman montrant les différences de valeurs seuils (Ct) dans deux essais TaqMan Hendra, M1 et N1, pour un échantillon témoin faiblement positif testé 28 fois.



Une mesure est faite avec chaque méthode sur chaque échantillon (modèle de paires appariées) pour comparer la répétabilité entre les deux méthodes sur la plage de mesure. Pour les tests de détection d'anticorps par ELISA compétitif, il est connu que la variabilité augmente avec la diminution de la concentration de l'analyte dans l'échantillon. Un témoin négatif ou faiblement positif peut avoir une variabilité significativement plus élevée que les témoins fortement positifs. La Figure 4 montre trois résultats qui se situent sur la ligne pour la valeur 0, à savoir la ligne où les résultats des deux tests sont identiques, leur différence étant nulle. Il s'agit des mêmes résultats que ceux tombant sur la ligne diagonale d'égalité sur le diagramme de dispersion de la Figure 2 pour les valeurs de 28,9, de 33,10 et de 33,45 dans les deux essais.

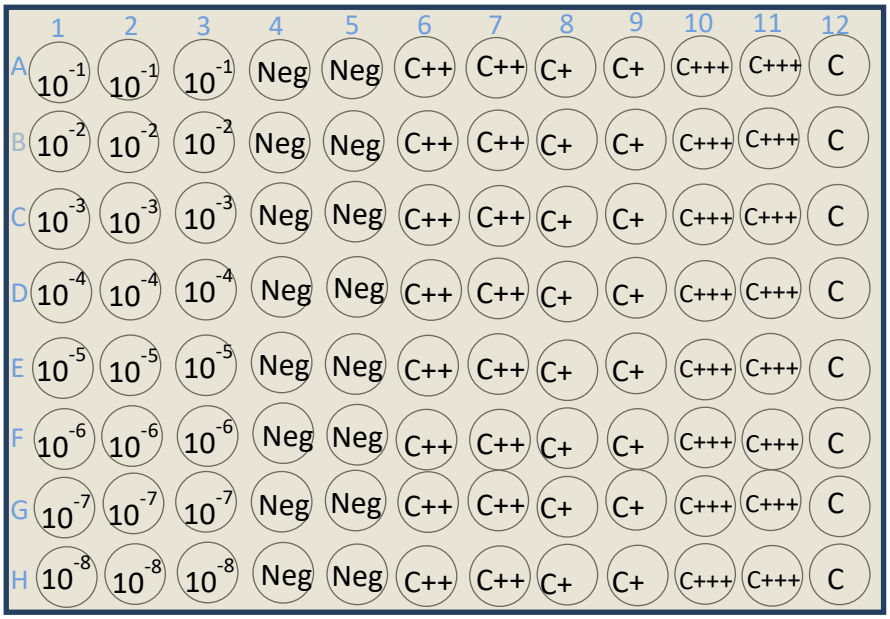
Tableau 4. Analyse statistique du graphique de Bland–Altman

Méthode A	TaqMan M1 faiblement positif
Méthode B	TaqMan N1 faiblement positif
Différences	
Taille de l'échantillonnage	28
Moyenne arithmétique	–1,8054
IC à 95 %	–2,1831 à –1,4276
Écart type	0,9741

E. EXPERIENCE POUR LA LIMITE DE DETECTION

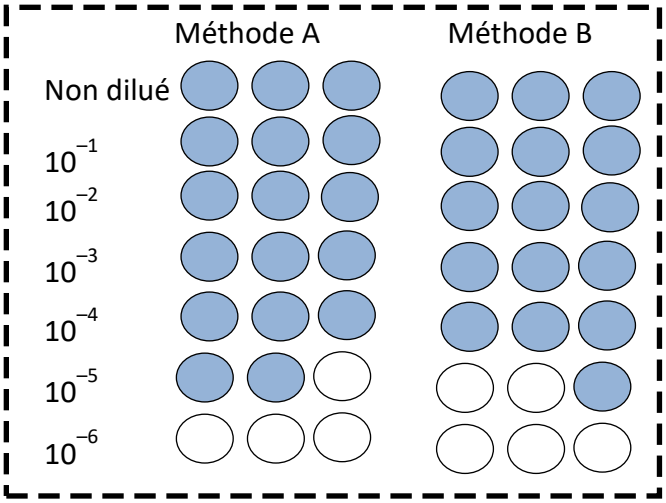
Un exemple de l'aspect d'une plaque pour une étude de comparabilité d'un test moléculaire est donné à la Figure 5. La limite de détection (trois copies d'échantillons positifs dans une dilution en série allant de 10^{-1} à 10^{-8} [sensibilité analytique]), la spécificité diagnostique (échantillons diagnostiques négatifs d'animaux non infectés ou d'animaux infectés avec un agent pathogène non cible [Neg] analysés en double), la sensibilité diagnostique (échantillons d'une activité variable provenant d'animaux infectés sur le terrain, soit extrêmement positifs [C+++], très fortement positifs [C++], fortement positifs [C+] et positifs [C], analysés en double) et la répétabilité ont été évaluées. Les contaminations croisées peuvent aussi être évaluées, les échantillons fortement positifs étant placés à côté des échantillons négatifs.

Fig. 5. Aspect d'une plaque 96 puits pour évaluer la Se analytique, la Se diagnostique et la Sp diagnostique ainsi que la répétabilité.



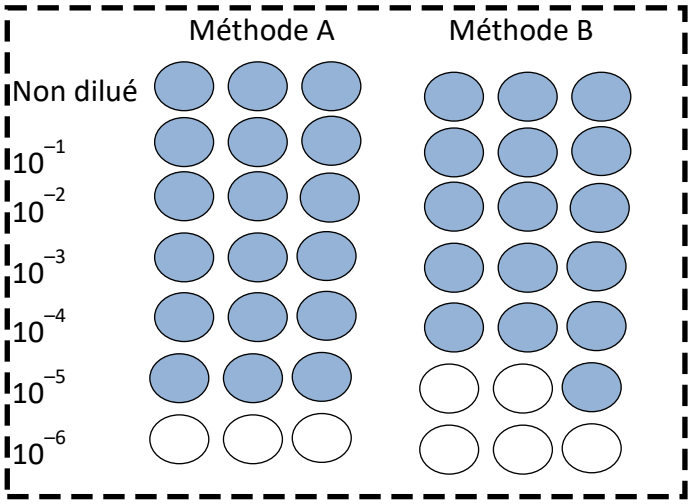
La limite de détection (LD) est une mesure de la sensibilité analytique (SeA) d'un essai. La LD est la quantité d'analyte dans une matrice donnée qui produirait un résultat positif au moins dans un certain nombre de cas (pourcentage). Les Figures 6 et 7 représentent les résultats hypothétiques d'une expérience de limite de détection. Par exemple, dans un titrage utilisant des dilutions décimales, toutes les copies à toutes les dilutions peuvent montrer une réponse soit de 100 %, soit de 0 %. Il y a deux options à ce stade. La dernière dilution montrant une réponse à 100 % peut être acceptée comme une estimation prudente de la limite inférieure de détection. Une estimation plus exacte peut être obtenue dans la phase suivante de l'expérience en utilisant des intervalles plus rapprochés dans le schéma de dilution et en se focalisant sur la zone entre 100 % et 0 %. La première étape consiste à produire, à répartir en aliquotes et à mettre en aveugle un nombre suffisant d'échantillons pour accomplir l'expérience et la seconde à produire un ensemble de dilutions de l'analyte, en utilisant de préférence la matrice de l'échantillon plutôt qu'un tampon comme diluant, reflétant la plage de mesure de la méthode, soit une série de dilutions décimales 10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}. Les exemples pratiques utilisent souvent 3-5 copies par étape de dilution. Utilisant une estimation prudente de 100 % (les 3 puits foncés à la dilution la plus élevée = positifs), les Figures 6 et 7 montrent respectivement des résultats comparables et des résultats non comparables. La méthode A représente la méthode validée et la méthode B la nouvelle méthode.

Fig. 6. Exemple d'une expérience de limite de détection (LD) avec un résultat acceptable.



Sur la Figure 6, toutes les répliques de la méthode A et de la méthode B à une dilution de 10^{-4} sont positives (bleu). À une dilution de 10^{-5}, seuls deux des trois puits sont positifs pour la méthode A et un seul des trois pour la méthode B. La limite de détection étant définie comme la dilution à laquelle tous les puits doivent être positifs, les résultats à partir de la dilution de 10^{-5} et au-dessous ne peuvent pas être pris en compte pour la comparabilité. En appliquant ces critères, les deux méthodes pouvant donc être considérées comme comparables, la limite de détection de la Figure 6 étant la même pour les méthodes A et B,.

Fig. 7. Exemple d'une expérience de limite de détection (LD) avec un résultat non acceptable.



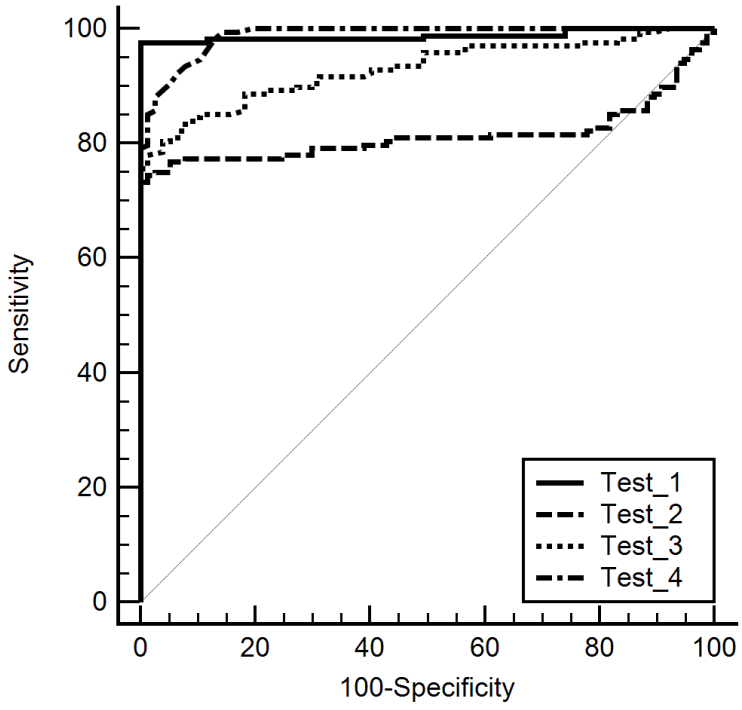
Sur la Figure 7, la dilution la plus élevée où les trois copies sont positives est de 10^{-5} pour la méthode A. Pour la méthode B en revanche, la dilution la plus élevée où les trois répliques sont toujours positives est de 10^{-4}. Par conséquent, les limites de détection de la méthode A et de la méthode B ne peuvent pas être considérées comme comparables dans la mesure où une différence d'une dilution logarithmique n'est pas acceptable. Il est conseillé de répéter l'expérience plusieurs fois avant de prendre une décision finale concernant la comparabilité des méthodes.

F. COMPARAISON DES COURBES ROC

L'analyse des caractéristiques récepteur-opérateur (ROC) est une méthode puissante pour évaluer et comparer l'exactitude globale d'un test diagnostique, c'est-à-dire la sensibilité diagnostique (SeD) et la spécificité diagnostique (SpD) à des seuils différents d'un ou de plusieurs tests modifiés (Greiner *et al.,* 2000). La mesure centrale est l'aire sous la courbe (AUC), la valeur de 1 indiquant un test avec une SeD de 100 % et une SpD de 100 %. Dans ce cas, il y a une séparation parfaite des valeurs des deux groupes, c'est-à-dire aucun chevauchement des distributions, et la courbe ROC atteint le coin supérieur gauche du graphique. En revanche, une valeur de 0,5 ne permet pas de faire une distinction entre les individus infectés et les individus non infectés au-delà de la probabilité ; la courbe ROC coïncide avec la diagonale, indiquant que le test est inutile. Les valeurs entre 0,5 et ≤0,7 peuvent être considérées comme peu précises, celles entre 0,7 et ≤0,9 comme moyennement précises et celles entre 0,9 et <1 comme très précises (Greiner *et al.,* 2000).

La Figure 8 montre les résultats de quatre titrages d'anticorps par ELISA pour le virus de l'influenza chez les porcs. Le panel était constitué de 168 sérums positifs et de 77 sérums négatifs, un test d'inhibition de l'hémagglutination ayant été utilisé comme test de référence (n = 245). Lorsque les titrages ont été comparés pour ces échantillons diagnostiques, le classement suivant a été établi : test 1 (AUC = 0,988), test 4 (AUC = 0,988), test 3 (AUC = 0,929) et test 2 (AUC = 0,814) (Tableau 5). Pour cette expérience basée sur un modèle de paires appariées, l'intervalle de confiance à 95 % pour les différences des aires sous les courbes ROC (AUC) peut être utilisé comme indicateur de la pertinence statistique (Tableau 6). En résumé, le test avec les meilleures SeD et SpD et avec l'aire sous la courbe la plus élevée (0,988) est le test 1. Ce résultat de 0,988 signifie qu'un individu choisi au hasard dans le groupe positif a une valeur de test plus élevée que celle d'un individu choisi au hasard dans le groupe négatif, et ce, dans 99 % des cas (Zweig et Campbell, 1993).

Fig. 8. Comparaison de l'aire sous la courbe (AUC) des caractéristiques récepteur-opérateur (courbes ROC) de quatre tests ELISA de détection des anticorps contre le virus de l'influenza chez les porcs.



Un autre moyen de comparer les résultats est d'évaluer les différences d'AUC. Ainsi, pour les tests 1 et 2, la différence entre les AUC était de 0,0173, pour les tests 1 et 3, de 0,058, pour les tests 1 et 4, de 0,00019, pour les tests 2 et 3, de 0,1150, pour les tests 2 et 4, de 0,174 et pour les tests 3 et 4, de 0,0586 (Tableau 6). Les tests avec les AUC les plus élevées et la différence d'AUC la plus faible étaient les tests 1 et 4, tous deux ayant une AUC de 0,988, chevauchant l'intervalle de confiance à 95 %, la différence entre les AUC ne dépassant pas 0,00019 et la valeur *p* de 0,98 indiquant l'absence de différence statistiquement significative à un niveau de 5 %. Les valeurs AUC basses, l'absence de chevauchement de l'IC à 95 %, les valeurs considérablement plus élevées des différences d'AUC et une valeur p < 0,05 indiquent l'absence de concordance des autres combinaisons de tests, à savoir 1 versus 3 ; 1 versus 2 ; 2 versus 3 ; 2 versus 4 et 3 versus 4. (test 1 : 0,964 à 0,997 ; test 2 : 0,760 à 0,861 ; test 3 : 0,889 à 0,958 ; test 4 : 0,965 à 0,997).

Des études comparatives plus complexes ayant recours à des tests basés sur des principes diagnostiques et biologiques analogues et utilisant des méthodes statistiques fréquentistes ou classiques ont été publiés (Brocchi *et al.,* 2006 ; Engel *et al.,* 2008). Voir aussi le Chapitre 2.2.5 *Méthodes statistiques de validation*.

Tableau 5. Comparaison de l'aire sous la courbe (AUC) des caractéristiques récepteur-opérateur (courbes ROC) et des valeurs* p *de quatre tests ELISA pour la détection des anticorps contre le virus de l'influenza chez les porcs.

Paramètre	**Test 1**	**Test 2**	**Test 3**	**Test 4**
Aire sous la courbe ROC	0,988	0,814	0,929	0,988

Tableau 6. Comparaison par paires des courbes ROC

	Test 1 vs 4	**Test 1 vs 3**	**Test 3 vs 4**	**Test 2 vs 3**	**Test 1 vs 2**	**Tetst 2 vs 4**
Différence entre les AUC	0,0002	0,055	0,0589	0,115	0,173	0,174
IC à 95 %	–0,016 to 0,016	0,025 to 0,092	0,027 to 0,090	0,069 to 0,161	0,117 to 0,230	0,118 to 0,229
Degré de signification	*p*=0,9808	*p* =0,0007	*p* =0,0003	*p* <0,0001	*p* <0,0001	*p* <0,0001

G. DISCUSSION ET CONCLUSIONS

Pour parvenir à la conclusion de la comparabilité ou non de deux méthodes, les résultats des expériences comparatives doivent être évalués au moyen d'analyses statistiques et d'évaluations objectives dans le but de contribuer à la prise de décision finale (NATA, 2013). D'autres critères tels que coûts/équipement, capacité de débit, temps d'exécution, capacités d'assurance qualité, sophistication technique, acceptation par la communauté régulatrice ou scientifique et compétences d'interprétation doivent cependant aussi être pris en compte lors du processus décisionnel (Figure 1, étape 6). Il est important de disposer d'une procédure établie, précisant qui a l'autorité ultime pour décider si les méthodes sont comparables ou non (responsable technique, directeur du laboratoire, responsable qualité).

Dans l'exemple ci-dessus, la comparaison de deux essais TaqMan indiquait une forte corrélation positive. La répétabilité des deux méthodes était comparable en appliquant un ET de 2-3 ou une valeur Ct de ± 2–3. Les valeurs Ct d'un échantillon témoin faiblement positif étaient constamment plus basses pour le TaqMan M1 que pour le N1, indiquant une Se légèrement supérieure de l'essai M1. Utilisé comme test de dépistage, le M1 serait plus adéquat, en raison de sa Se supérieure. Ces résultats ont été corroborés par l'analyse des échantillons diagnostiques de foyers de virus Hendra entre 2011 et 2013 (données non incluses). Par ailleurs, disposer de deux tests ciblant des gènes différents augmente la chance de ne pas manquer un nouveau variant. Dans le cas d'une zoonose mortelle comme le virus Hendra, il s'agit d'une considération importante.

Le graphique de Bland-Altman dans la Figure 4 et le Tableau 4 montre que les résultats statistiques sont parfois difficiles à interpréter, notamment parce que l'intervalle de confiance à 95 % pour la différence moyenne exclut zéro. Cela peut être interprété comme un défaut de comparabilité, mais la modification des valeurs seuils peut servir à compenser ces résultats.

L'analyse ROC des 4 ELISA de l'influenza porcine (Figure 8 et Tableau 5 et 6) indique que deux des tests ont une SeD et une SpD quasiment identiques (tests 1 et 4). En même temps, cela aide à classer la performance des autres tests. Dans ces circonstances, le coût, la disponibilité et d'autres critères détermineront la décision finale quant au test le plus approprié pour un objectif donné.

Par ailleurs, il n'est pas nécessaire d'évaluer tous les paramètres dans chaque étude comparative des méthodes. Par exemple, une évaluation de la contamination croisée sera nécessaire pour des changements d'équipement, tel le passage d'une procédure manuelle à une procédure robotisée pour l'extraction de l'acide nucléique, mais ne le sera pas pour le changement d'un réactif important, telle une amorce ou une sonde différente. Il est de règle de décider des paramètres pertinents et des critères d'acceptation ou de rejet avant de commencer l'expérience, la question la plus importante étant l'adéquation du nouveau test à l'objectif.

Les analyses diagnostiques vétérinaires sont de nature à garder une certaine flexibilité lorsqu'il s'agit de spécifier les limites d'acceptation. Ainsi, lors de la comparaison de deux tests moléculaires, il peut arriver que la différence ne dépasse pas 1, 2 ou 3 Ct pour 95 % (99 %) des échantillons testés, 10 % de la moyenne de deux échantillons pour au moins 95 % (99 %) des échantillons testés, ni 1, 2 ou 3 ET des échantillons. Il est recommandé de prendre en compte les limites et les paramètres importants dès le début, les considérations qualitatives (coût, simplicité de l'analyse, rapidité des résultats) intervenant probablement en dernier. Quel que soit le paramètre choisi comme règle de base, le test candidat ne devrait pas, de manière générale, fonctionner significativement moins bien que le test validé.

Les données générées et le processus de prise de décision pour l'acceptabilité des changements doivent être clairement documentés et conservés pour pouvoir servir de piste d'audit.

H. ANALYSE DES DONNEES

Les données du présent document ont été stockées et regroupées dans Microsoft Excel. L'analyse et la reproduction graphique des diagrammes de dispersion et des histogrammes, la distribution des données et leur reproduction graphique, les graphiques de Bland-Altman, les graphiques comparatifs et les analyses ROC ont été effectués à l'aide de MedCalc (MedCalc®, version 12.4.0.0, 64 bit, Window XP/Vista 7/8, *www.medcalc.org*, Copyright 1993–2013, MedCalc software bvba).

RÉFÉRENCES

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

BLAND J.M. & ALTMAN D.G. (2007). Agreement between methods of measurement with multiple observations per individual. *J. Biopharm. Stat.*, **17**, 571–582.

BROCCHI E., BERGMANN I.E., DEKKER A. PATON D.J., SAMMIN D.J., GREINER M., GRAZIOLI S., DE SIMONE F., YADIN H., HAAS B., BULUT N., MALIRAT V., NEITZERT E., GORIS N., PARIDA S., SØRENSEN K. & DE CLERCQ K. (2006). Comparative evaluation of six ELISAs for the detection of antibodies to the non-structural proteins of foot-and-mouth disease virus. *Vaccine*, **24**, 6966–6979.

ENGEL B., BUIST W., ORSEL K., DEKKER A., DE CLERCQ C., GRAZIOLI S. & VAN ROERMUND H. (2008). A Bayesian evaluation of six diagnostic tests for food-and-mouth disease for vaccinated and non-vaccinated cattle. *Prev. Vet. Med.*, **86**, 124–138.

GALL D., COLLING A., MARINO O., MORENO E., NIELSEN K., PEREZ B. & SAMARTINO L. (1998). Enzyme immunoassays for serological diagnosis of bovine brucellosis: a trial in Latin America. *Clin. Diagn. Lab. Immunol.*, **5**, 654–651.

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.

KOZAK M. & WNUK A. (2014). Including the Tukey mean-difference (Bland–Altman) plot in a statistics course. *Teaching Statistics*, **36**, 83–87.

NATIONAL ASSOCIATION OF TESTING AUTHORITIES (OF AUSTRALIA) (NATA) (2013). NATA Technical Note 17, October 2013. Guidelines for the validation and verification of quantitative and qualitative test methods. http://www.nata.com.au/nata/phocadownload/publications/Guidance_information/tech-notes-information-papers/technical_note_17.pdf (accessed 8 April 2015)

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 2016.