Guide pratique de validation statistique de méthodes de mesure : répétabilité, reproductibilité, et concordance

Loic Desquilbet

Guide pratique de validation statistique de méthodes de mesure : répétabilité, reproductibilité, et concordance

Loïc Desquilbet, PhD en Santé Publique

Professeur en Biostatistique et en Epidémiologie Clinique
Département des Sciences Biologiques et Pharmaceutiques
Ecole nationale vétérinaire d'Alfort

Version 12.8 (28/01/2020)

Préface

Résumé

Une méthode de mesure d'un caractère, qu'elle soit objective ou subjective, devrait au préalable avoir prouvé sa validité avant d'être utilisée. Cette étape de validation est indispensable lorsque cette méthode de mesure est utilisée pour collecter des données qui serviront à la recherche scientifique ; elle l'est tout autant si cette méthode est utilisée dans le cadre du suivi médical d'un patient. La *répétabilité* et la *reproductibilité* d'une méthode de mesure d'un caractère font partie des critères de validation de cette méthode de mesure. De plus, si cette méthode de mesure vise à remplacer une autre considérée comme une méthode de référence, un critère de validation supplémentaire est celui de la *concordance* des valeurs du caractère fournies par les deux méthodes. Ce document est un guide pratique pour quantifier, à l'aide d'indicateurs statistiques simples et publiés dans la littérature scientifique, la répétabilité et la reproductibilité d'une méthode de mesure, et la concordance entre deux méthodes de mesure. Des exemples issus de la médecine vétérinaire permettent d'illustrer la quantification de la répétabilité, reproductibilité, et concordance, que le caractère mesuré soit binaire, qualitatif, ou bien quantitatif. Le document présente enfin le protocole à mettre en place ainsi que le nombre d'individus à mesurer afin d'estimer les indicateurs statistiques de répétabilité, reproductibilité, et de concordance avec suffisamment de précision pour garantir la validation de la méthode de mesure.

Public cible

Toute personne impliquée dans la collecte de données dont elle souhaite montrer la validité de la méthode de mesure de ces données. Des référentiels réglementaires de validation de méthode de mesure existant pour les laboratoires d'analyses biologiques, ce document ne vise donc pas les personnes travaillant dans ces laboratoires d'analyses. A part ces personnes là, ce document peut intéresser des personnes de disciplines très différentes (médecines humaine et vétérinaire, sociologie, éthologie, sciences de l'ingénieur, …).

Remerciements

Je tiens à remercier les Dr vétérinaires C Boyer et M Huynh de m'avoir permis d'utiliser la radio de la figure 7, et de m'avoir fourni les données issues de la thèse vétérinaire de C Boyer.

Contrat de diffusion

Pas de modifications — Dans le cas où vous effectuez un remix, que vous transformez, ou créez à partir du matériel composant l'Oeuvre originale, vous n'êtes pas autorisé à distribuer ou mettre à disposition l'Oeuvre modifiée.

Citation du document

Vous pouvez citer ce document de la façon suivante : Desquilbet, L. 2019. *Guide pratique de validation statistique de méthodes de mesure : répétabilité, reproductibilité, et concordance*. [En ligne, disponible à : https://hal.archives-ouvertes.fr/hal-02103716]

Questionnaire de satisfaction en ligne

Afin de recueillir votre avis sur ce document, et de l'améliorer, j'ai créé un questionnaire de satisfaction totalement anonyme qui ne vous prendra pas plus de 5 minutes à remplir (9 questions), en cliquant sur le lien suivant : https://goo.gl/forms/1jqv4PlndVGytXEz1

Je vous remercie beaucoup par avance de bien vouloir remplir ce questionnaire. Par ailleurs, si vous voyez des coquilles dans le texte, n'hésitez pas à me les signaler par email (loic.desquilbet(at)vet-alfort.fr) !

Table des matières

I. Introduction

A. Contexte et objectifs

Le contexte de l'évaluation de la répétabilité et de la reproductibilité d'une méthode de mesure, ou de la concordance de méthodes de mesure est celui de la Démarche Qualité : nous souhaitons évaluer la *qualité* d'une méthode de mesure avant d'utiliser cette méthode à des fins scientifiques, médicales, ou autres. Il peut s'agir soit d'une nouvelle méthode de mesure, soit d'une méthode de mesure déjà existante et utilisée, mais dont la qualité n'a pas été encore évaluée. L'objectif visé est d'avoir confiance dans les valeurs fournies par la méthode de mesure.

B. Définitions de termes employés

1. Notions de « Caractère » et de « valeur du caractère »

Le « caractère » (appelé « mesurande » en métrologie) est la grandeur soumise à la mesure (« mesurage » en métrologie). Il existe quatre types de caractères :

- Caractère binaire : le caractère est présent ou absent (par exemple, la présence/absence d'une tumeur) ;
- Caractère qualitatif nominal : le caractère est sous forme de classes non ordonnées (par exemple, le type de la tumeur) ;
- Caractère qualitatif ordinal : le caractère est sous forme de classes ordonnées (par exemple, le grade de la tumeur) ;
- Caractère quantitatif : le caractère est quantifiable par une valeur numérique, avec ou sans chiffre après la virgule (par exemple, la taille de la tumeur).

Les valeurs du caractère sont « présence / absence » pour un caractère binaire, la classe du caractère pour un caractère qualitatif, et la valeur du caractère pour un caractère quantitatif.

2. Notion de « méthode de mesure »

La « méthode de mesure » doit être prise dans son sens le plus large possible. Parmi les méthodes de mesure les plus fréquemment rencontrées, je peux citer les suivantes : l'évaluation subjective du caractère à l'aide de l'un des cinq sens (vue, ouïe, odorat, toucher, et goût), à l'aide d'un questionnaire, à l'aide d'un instrument de mesure « mécanique » (capteurs, règles, …) ou « informatique » (un programme informatique mesure le caractère à l'aide d'algorithmes divers). Cette liste n'est bien entendu pas exhaustive.

3. Notion d' « individu »

Dans tout ce guide pratique, je vais entendre par « individu » une « entité » indépendante d'une autre « entité » sur laquelle les mesures vont être réalisées (mesures réalisées soit par le même opérateur, soit par deux opérateurs différents). Par exemple, un « individu » peut être un animal, une radiographie, une coupe histologique, un prélèvement biologique, *etc*.

4. *Fidélité, répétabilité, reproductibilité (définitions ISO)*

Les définitions ci-dessous sont issues du document « Vocabulaire international de métrologie – Concepts fondamentaux et généraux et termes associés (VIM) », téléchargeable ici[1].

La **fidélité** est « l'étroitesse de l'accord entre les indications ou les valeurs mesurées obtenues par des mesurages répétés du même objet ou d'objets similaires dans des conditions spécifiées. »

La **répétabilité** est la fidélité dans les conditions de mesures suivantes : « conditions qui comprennent la même procédure de mesure, les mêmes opérateurs, le même système de mesure, les mêmes conditions de fonctionnement et le même lieu, ainsi que des mesurages répétés sur le même objet ou des objets similaires pendant une courte période de temps. »

La **reproductibilité** est la fidélité dans les conditions de mesures suivantes : « conditions qui comprennent des lieux, des opérateurs et des systèmes de mesure différents, ainsi que des mesurages répétés sur le même objet ou des objets similaires. »

5. *Commentaires sur ces définitions*

Une méthode de mesure répétable ou reproductible n'est pas forcément une méthode qui fournit des mesures *correctes*. Une méthode de mesure est valide si elle est répétable, reproductible, *et si* elle fournit des valeurs correctes. Pour cela, il faudra confronter les valeurs fournies par la méthode de mesure testée à celles fournies par une méthode dite de référence.

C. Contextes de répétabilité, de reproductibilité, et de concordance

Le tableau 1 résume les différents contextes de mesures lorsque l'on souhaite quantifier la répétabilité, la reproductibilité, et la concordance de méthode(s) de mesure.

Tableau 1. Récapitulatif des contextes de mesures pour quantifier la répétabilité, la reproductibilité, et la concordance de méthode(s) de mesure.

	Opérateur(s)		Conditions de mesures		Méthodes de mesure	
Contexte n°	Identique	Différents	Identique	Différentes	Identiques	Différentes
1	X		X		X	
2		X	X		X	
3	X			X	X	
4	X	(X)	X	(X)		X

Le contexte n°1 correspond à la situation où l'on souhaite savoir si une *même* méthode de mesure, utilisée par un *même* opérateur, est reproductible lorsque les conditions de mesures *ne varient quasiment pas* (lorsque les mesures sont espacées par un intervalle de temps jugé *a priori* négligeable). On parlera alors de **répétabilité** de la méthode de mesure (Barnhart et al., 2007). Si les conditions de mesures varient alors même que l'intervalle de temps entre les mesures est jugé comme faible, on ne parlera pas de répétabilité, mais de reproductibilité (spatio-temporelle).

Le contexte n°2 correspond à la situation où l'on souhaite savoir si deux opérateurs, lorsqu'ils évaluent la *même* chose avec la *même* méthode de mesure, donnent un *même* avis (si le critère est binaire ou qualitatif à ≥ 3 classes) ou quantifient de la *même* quantité ce qu'ils viennent d'évaluer. On parlera dans ce cas-là de **reproductibilité inter-opérateurs** de la méthode de mesure. A noter que, dans ce contexte, les conditions de mesures doivent être identiques. Si elles ne le sont pas en pratique, il faudra alors considérer que les différences de conditions de mesures ne vont pas avoir d'impact dans l'évaluation de la reproductibilité inter-opérateurs de la méthode de mesure.

[1] https://www.bipm.org/en/publications/guides/

Le contexte n°3 correspond à la situation où l'on souhaite savoir si une *même* méthode de mesure, utilisée par un *même* opérateur, est reproductible lorsque les conditions de mesures *varient* (ou bien lorsque les mesures sont espacées par un intervalle de temps jugé *a priori* non négligeable). On parlera dans ce cas-là de **reproductibilité spatio-temporelle** de la méthode de mesure. A noter que, dans ce contexte, il ne devrait y avoir qu'un seul opérateur. Si ce n'est pas le cas en pratique, il faudra alors considérer que la reproductibilité inter-opérateurs est excellente.

Le contexte n°4 correspond à la situation où l'on souhaite savoir si *deux* méthodes de mesure fournissent les *mêmes* résultats lorsqu'elles mesurent la *même* chose dans les *mêmes* conditions. On parlera alors de **concordance entre deux méthodes de mesure**. Dans cette situation, il n'y a en général qu'un seul opérateur. Mais si l'on peut considérer qu'il n'y a pas d' « effet » opérateur, alors cette situation peut s'appliquer à la situation où les opérateurs sont différents. A noter que, là encore dans cette situation, on préfèrera que les conditions de mesures des deux méthodes soient identiques, mais en pratique, elles peuvent ne pas l'être. Si elles ne le sont effectivement pas, il faudra alors considérer que les différences de conditions de mesures ne vont pas avoir d'impact dans l'évaluation de la concordance entre les deux méthodes de mesure. La concordance de deux méthodes de mesure est indispensable à évaluer lorsque l'on souhaite savoir si la nouvelle méthode fournit des valeurs correctes (l'une des deux méthodes sera donc celle considérée comme de référence).

D. Notion de « série de mesures »

Une « série de mesures » est un ensemble de *N* mesures réalisées *une* fois sur *N* individus, par un même opérateur, dans des mêmes conditions de mesures, avec la même méthode de mesure. Pour quantifier la répétabilité / reproductibilité d'une méthode de mesure, ou la concordance de deux méthodes de mesure, les indicateurs numériques et graphiques que l'on va voir dans ce guide pratique vont nécessiter *deux* séries de mesures : une série par opérateur avec deux opérateurs (reproductibilité inter-opérateurs), une série par condition de mesures avec deux conditions différentes (répétabilité ou reproductibilité spatio-temporelle), ou bien une série par méthode de mesure avec deux méthodes différentes (concordance de méthodes de mesure). Un nombre de séries de mesures supérieur ou égal à trois est possible, mais nécessite un traitement statistique beaucoup plus poussé, sans que l'apport soit majeur (Giraudeau and Mary, 2001).

Si la différence entre les contextes n°1 et n°3 (c'est-à-dire la différence entre la répétabilité d'une méthode de mesure et la reproductibilité spatio-temporelle) est uniquement l'intervalle de temps entre les deux séries de mesures sur le même individu, et s'il est (jugé comme) très court, on parlera de répétabilité ; s'il est jugé comme suffisamment long pour avoir potentiellement un impact, on parlera de reproductibilité (spatio-temporelle). On peut le voir aussi de façon inverse : pour quantifier la répétabilité d'une méthode de mesure, l'intervalle de temps doit être (jugé comme) très court, c'est-à-dire trop court pour penser que cet intervalle de temps puisse avoir un impact ; pour quantifier la reproductibilité spacio-temporelle d'une méthode de mesure, l'intervalle de temps ou la différence d'espace doit être celui/celle dont on veut tester la reproductibilité (par exemple, « est-ce que ma méthode de mesure va me donner le même résultat si j'évalue ce que je veux évaluer à une semaine d'intervalle ? » ou bien « est-ce que ma méthode de mesure va me donner le même résultat si j'évalue ce que je veux évaluer à deux endroits ou conditions différent(e)s ? »).

Dans toute la suite de ce guide pratique, je ne parlerai plus que de « concordance » de deux séries de mesures. Les différents contextes de mesure de ces deux séries de mesures présentés dans le tableau 1 vont conduire au fait que nous allons quantifier la *concordance* de deux séries de mesures pour évaluer soit la répétabilité, soit la reproductibilité inter-opérateurs, soit de reproductibilité spatio-temporelle d'une même méthode de mesure, soit enfin la concordance entre deux méthodes de mesure.

Le tableau 2 ci-dessous présente la structure d'un fichier de données nécessaire pour estimer les indicateurs statistiques décrits dans ce guide pratique.

Tableau 2. Structure d'un fichier de données nécessaire pour estimer la concordance de deux séries de mesures.

N° de l'individu	Série n°1	Série n°2
1	Valeur 1-1	Valeur 1-2
2	Valeur 2-1	Valeur 2-2
3	Valeur 3-1	Valeur 3-2
4	Valeur 4-1	Valeur 4-2
5	Valeur 5-1	Valeur 5-2
…	Valeur …-1	Valeur …-2

Dans le tableau ci-dessus, par exemple, « valeur 4-2 » est la valeur du caractère de l'individu n°4 lors de sa 2ème mesure.

E. Petite revue de la littérature

Différentes méthodes existent pour évaluer la répétabilité, reproductibilité, ou concordance de méthode(s) de mesure, et sont présentées dans deux excellents articles (Patton et al., 2006; Watson and Petrie, 2010). Je présenterai dans ce guide pratique le coefficient de concordance Kappa lorsque le caractère est binaire ou qualitatif, et le coefficient de concordance de Lin ainsi que le graphique de Bland et Altman lorsque le caractère est quantitatif. Ce sont en effet les principaux indicateurs de répétabilité/reproductibilité/concordance de deux séries de mesures (Lin et al., 2007; Ludbrook, 2002), largement utilisés en recherche clinique vétérinaire (Bergknut et al., 2013; Durando et al., 2008; Gibbons-Burgener et al., 2001; Giori et al., 2011; Norton et al., 2011; Perkins et al., 2009; Tennent-Brown et al., 2011; Voyvoda and Erdogan, 2010).

F. Ressources Excel®

Le guide pratique va utiliser deux fichiers Excel® qui peuvent vous être fournis par email en m'envoyant votre demande à loic.desquilbet(at)vet-alfort.fr. Ils permettent de calculer les deux coefficients de concordance cités dans ce guide pratique (Kappa et Lin) et de dresser le graphique de Bland et Altman. Toutes les copies d'écran d'Excel® dans ce document proviennent de ces deux fichiers Excel®.

II. Concordance entre deux séries de mesures binaires

A. Présentation d'un exemple

Supposons que pour savoir si une vache est atteinte de réticulo-péritonite traumatique (RPT), on appuie fortement sur le thorax de la vache, et si la vache exprime de la douleur, le vétérinaire va conclure qu'il y a présence de RPT. Dans cet exemple, on peut imaginer que, pour une même vache (qu'elle ait ou non réellement une RPT), deux vétérinaires peuvent conclure différemment quant à la présence ou l'absence de RPT (l'évaluation de la douleur de la vache peut être considérée comme subjective). Supposons une étude dont l'objectif est de quantifier la reproductibilité inter-opérateurs de la méthode de diagnostic de RPT suivante : « appuyer fortement sur le thorax de la vache et si la vache exprime de la douleur, c'est que la vache présente une RPT ». Le protocole de cette étude est le suivant : 64 vaches ont été examinées par deux vétérinaires qui ont tous les deux appliqué la méthode de diagnostic testée. Le tableau 3 présente le nombre de diagnostics de présence de RPT et le nombre de diagnostics d'absence de RPT posés par les deux vétérinaires sur les mêmes 64 vaches.

Tableau 3. Répartition des diagnostics de réticulo-péritonite traumatique (RPT) réalisés par 2 vétérinaires sur 64 vaches examinées.

		Diagnostic vét. n°2		
		RPT absente	RPT présente	Total
Diagnostic vét. n°1	RPT absente	17	4	21
	RPT présente	3	40	43
	Total	20	44	64

B. Calcul du coefficient de concordance Kappa

L'accord observé entre les deux vétérinaires jugeant la présence (ou l'absence) de la RPT à partir des 64 vaches qu'ils auront tous les deux examinées résulte de la somme d'une composante d'accord « aléatoire » (accord dû simplement au hasard) et d'une composante d'accord « véritable ». Cette notion d'accord « aléatoire » est importante à saisir : si les deux vétérinaires évaluaient *au hasard* la présence de douleurs (c'est-à-dire ici, en jugeant à pile ou face le fait que la vache exprime de la douleur), alors il y aurait un nombre non nul de vaches pour lesquelles les deux vétérinaires diraient tous les deux « absence de RPT » et « présence de RPT ».

Pour prendre en compte le phénomène d'accord aléatoire, le coefficient de concordance Kappa (K) quantifie l'intensité de l'accord « véritable » (Cohen, 1960; Kraemer et al., 2002; Sim and Wright, 2005). C'est un indice qui vise à « enlever » la portion de hasard dans l'accord observé entre les deux vétérinaires.

Le tableau 3 montre que la méthode de diagnostic de RPT donne des diagnostics concordants pour 57 (89%) vaches examinées (17 vaches avec RPT absente et 40 vaches avec RPT présente par les deux vétérinaires), et discordants pour 7 (11%) vaches.

La formule du coefficient de concordance Kappa est la suivante :

$$K = \frac{C_{obs} - C_{al}}{1 - C_{al}}$$

Avec :

C_{obs} = concordance observée

C_{al} = concordance aléatoire

La concordance observée est une proportion qui vaut la somme des effectifs concordants observés divisée par la taille de l'échantillon total. Ici, C_{obs} = (17+40)/64 = 0,89.

La concordance aléatoire est une proportion qui vaut la somme des effectifs concordants théoriques divisée par la taille de l'échantillon total. Pour calculer cette concordance aléatoire, il faut d'abord calculer les effectifs que l'on aurait observés dans chacune des deux cases concordantes si les deux vétérinaires avaient posé leur diagnostic complètement au hasard (⇔ « effectifs concordants théoriques »). Ces effectifs sont calculés de la même façon que dans le calcul d'un test du Chi-2 : le produit des deux marges divisé par la taille totale de l'échantillon. Ici, C_{al} = [(20*21)/64+(44*43)/64] / 64 = 0,56

D'où, K = +0,75.

La figure 1 ci-dessous est une copie d'écran du fichier Excel® fournissant la valeur du coefficient de concordance Kappa et son intervalle de confiance à 95% (en utilisant la méthode « goodness-of-fit » présentée dans l'article de Donner et Eliasziw (Donner and Eliasziw, 1992)) : 0,75 [0,53 ; 0,88].

		Présence caractère série 2		
		Absent	Présent	Total
Présence caractère série 1	Absent	17	4	21
	Présent	3	40	43
	Total	20	44	64

Kappa [IC95%]	0,75 [0,53 - 0,88]
P-value	< 0,001

Interprétation de la valeur du Kappa (Landis et Koch, Biometrics, 1977)	
< 0	Très mauvais (Poor)
0-0.20	Mauvais (Slight)
0.21-0.40	Passable (Fair)
0.41-0.60	Moyenne (Moderate)
0.61-0.80	Bonne (Substantial)
0.81-1.00	Très bonne (Almost perfect)

Figure 1. Valeur du coefficient de concordance Kappa et son intervalle de confiance à 95% dans l'exemple de diagnostics de réticulo-péritonite traumatique (RPT) réalisés par 2 vétérinaires sur 64 vaches examinées.

C. Interprétation de la valeur du coefficient de concordance Kappa

Le coefficient de concordance Kappa est un nombre sans dimension compris entre -1 et +1. L'accord est d'autant plus élevé que la valeur du coefficient Kappa est proche de +1. Une valeur du coefficient Kappa est égale à -1 lorsqu'il n'y a aucune réponse concordante entre les deux vétérinaires (désaccord parfait). Lorsqu'il y a indépendance des jugements, le coefficient Kappa est égal à zéro ($C_{obs} = C_{al}$).

Suivant le classement de Landis et Koch présenté dans le tableau 4 (Landis and Koch, 1977), et qui est fréquemment utilisé en biologie, le coefficient Kappa de l'exemple (K=+0,75) aurait conduit à penser que la reproductibilité inter-opérateurs de la méthode de diagnostic de RPT « appuyer fortement sur le thorax de la vache et si la vache exprime de la douleur, c'est que la vache présente une RPT » est « bonne ».

Tableau 4. Interprétation des valeurs du coefficient Kappa (Landis and Koch, 1977).

Coefficient Kappa	Interprétation
< 0	Très mauvais (Poor)
0-0,20	Mauvais (Slight)
0,21-0,40	Passable (Fair)
0,41-0,60	Moyenne (Moderate)
0,61-0,80	Bonne (Substantial)
0,81-1,00	Très bonne (Almost perfect)

D. Danger d'une sur-interprétation du coefficient Kappa

Veuillez vous rappeler (cf. partie I.B.5) que ce n'est pas parce les deux vétérinaires ont posé le même diagnostic sur 57 vaches parmi les 64 que ces 57 diagnostics sont *corrects*. Par conséquent, dans le contexte de reproductibilité inter-opérateurs d'une méthode, un excellent coefficient de concordance n'est pas forcément synonyme d' « excellente méthode » ! Pour s'assurer que la méthode de mesure fournit des valeurs correctes, il est indispensable de calculer la concordance entre la méthode de mesure évaluée et une méthode de mesure de référence. Dans ce contexte, en plus de la valeur du coefficient Kappa, les valeurs de sensibilité (Se) et spécificité (Sp) pourront être aussi calculées, ce qui permettra une meilleure interprétation du coefficient de concordance (cf. ci-dessous).

E. Coefficient Kappa, sensibilité, et spécificité dans l'évaluation de la concordance de méthodes de mesure

Dans le contexte de savoir si une nouvelle méthode donne des valeurs identiques à une méthode de référence avec un caractère binaire (présent/absent), rappelons que la sensibilité de la nouvelle méthode sera d'autant plus importante qu'elle identifiera peu de faux négatifs (FN ; cf. tableau 5) parmi tous les individus vraiment positifs (TP). La spécificité de la nouvelle méthode sera d'autant plus importante qu'elle identifiera peu de faux positifs (FP ; cf. tableau 5) parmi tous les individus vraiment négatifs (TN). Autrement dit, la nouvelle méthode sera dite « sensible » si elle est grandement capable d'identifier les individus comme « positifs » parmi les individus vraiment positifs. La nouvelle méthode sera dite « spécifique » si elle est grandement capable d'identifier les individus comme « négatifs » parmi les individus vraiment négatifs.

Tableau 5. Présentation des données pour des calculs de sensibilité et spécificité.

		Méthode de référence	
		Caractère absent (-)	Caractère présent (+)
Nouvelle méthode	Caractère absent (-)	VN	FN
	Caractère présent (+)	FP	VP
	Total	TN	TP

Les formules pour calculer les sensibilité (Se) et spécificité (Sp) sont les suivantes :

$$Se = \frac{VP}{VP + FN} = \frac{VP}{TP}$$

$$Sp = \frac{VN}{VN + FP} = \frac{VN}{TN}$$

On peut se rendre compte que les Se et Sp n'utilisent, chacune, qu'une partie de l'échantillon (les TP individus vraiment positifs pour la Se, et les TN individus vraiment négatifs pour la Sp), alors que le coefficient Kappa, qui peut se calculer à partir du même tableau 5, va utiliser l'ensemble de l'échantillon puisque son calcul va utiliser VN, VP, et les marges du tableau 5 (cf. partie II.B). Par conséquent, le coefficient Kappa peut être vu comme une « moyenne » de la Se et de la Sp dans le contexte de concordance entre deux méthodes de mesure, l'une des deux méthodes étant la méthode de référence. Ainsi, une faible valeur du coefficient Kappa pourra être expliquée grâce au calcul de la Se et de la Sp : le coefficient de concordance Kappa est faible à cause d'une faible valeur de Se, de Sp, ou des deux ? La réponse à cette question permettra de savoir quelle partie de la nouvelle méthode de mesure il faut améliorer lorsque le coefficient de concordance avec la méthode de référence est < 0,60 : améliorer sa sensibilité, sa spécificité, ou bien les deux.

III. Concordance entre deux séries de mesures qualitatives

A. Caractère qualitatif nominal

Le calcul de coefficient de concordance Kappa classique que nous avons vu dans la partie II s'étend tout à fait au cas d'un caractère qualitatif nominal (Cohen, 1960).

Prenons l'exemple du tableau clinique évoquant la fièvre aphteuse chez la vache, avec deux vétérinaires posant leur diagnostic sur 56 vaches. Les tableaux 6.1 et 6.2 présentent deux situations différentes (cf. valeurs en gras et soulignées). La valeur du coefficient de concordance Kappa est pourtant identique entre la situation n°1 (cf. tableau 6.1) où le vétérinaire 1 pose le diagnostic « maladie des muqueuses » sur 7 vaches et le vétérinaire 2 pose le diagnostic « fièvre aphteuse » pour ces 7 mêmes vaches (coefficient kappa = 0,59), et la situation n°2 (cf. tableau 6.2) où le vétérinaire 1 pose le diagnostic « autre diagnostic » sur 7 vaches et le vétérinaire 2 pose le diagnostic « fièvre aphteuse » pour ces 7 mêmes vaches (coefficient kappa = 0,59).

Tableau 6.1. Diagnostics réalisés par deux vétérinaires sur des 56 vaches présentant des signes cliniques évoquant la fièvre aphteuse (situation n°1).

		Diagnostic vét. 2				Total
		fièvre aphteuse	maladie muqueuses	coryza gangréneux	Autre diagnostic	
Diagnostic vét. 1	fièvre aphteuse	12	3	0	1	16
	maladie muqueuses	**7**	8	1	0	16
	coryza gangréneux	0	3	6	0	9
	Autre diagnostic	**0**	0	2	13	15
	Total	19	14	9	14	56

Tableau 6.2. Diagnostics réalisés par deux vétérinaires sur des 56 vaches présentant des signes cliniques évoquant la fièvre aphteuse (situation n°2).

		Diagnostic vét. 2				Total
		fièvre aphteuse	maladie muqueuses	coryza gangréneux	Autre diagnostic	
Diagnostic vét. 1	fièvre aphteuse	12	3	0	1	16
	maladie muqueuses	**0**	8	1	0	9
	coryza gangréneux	0	3	6	0	9
	Autre diagnostic	**7**	0	2	13	22
	Total	19	14	9	14	56

L'égalité de ces deux valeurs du coefficient Kappa (0,59 dans les deux situations) est attendue, puisque ces deux situations révèlent une même absence de concordance : dans les deux situations, 7 vaches ne sont pas classées de la même façon par les deux vétérinaires. En pratique, il peut arriver que les deux valeurs du coefficient Kappa ne soient pas exactement identiques en intervertissant les effectifs discordants. Ce n'est pas une erreur : n'oubliez pas (cf. partie II.B) que le calcul du coefficient Kappa utilise non seulement les paires concordantes situées sur la diagonale du tableau, mais aussi les marges (pour calculer la concordance aléatoire). En intervertissant les effectifs dans les paires discordantes, les effectifs concordants ne sont certes pas modifiés, mais les marges, elles, le sont, ce qui conduit à des valeurs du coefficient Kappa modifiées (à la marge).

Il faut noter que des limites ont été décrites par Maclure et Willett dans l'interprétation du coefficient Kappa dans le cas de variables qualitatives nominales (Maclure and Willett, 1987). Les deux auteurs recommandent de regrouper les catégories de telle sorte à n'avoir que des caractères binaires (par exemple : fièvre aphteuse *versus* pas de fièvre aphteuse) et calculer autant de coefficients Kappa que de nouveaux caractères binaires ainsi créés.

B. Caractère qualitatif ordinal

Le calcul de coefficient de concordance Kappa classique peut aussi s'étendre au cas des caractères qualitatifs ordinaux. Cependant, si l'on calcule un coefficient de concordance Kappa de façon classique (comme cela vient d'être fait avec l'exemple de la fièvre aphteuse) avec un caractère qualitatif *ordinal*, ce caractère sera considéré comme qualitatif *nominal* dans les calculs, et l'interprétation devra en tenir compte ! L'exemple qui suit va le montrer.

Supposons le caractère qualitatif ordinal « degré de sédation après prémédication » en 4 classes : « sédation absente ou quasi absente », « sédation légère », « sédation modérée », et « sédation importante ». Supposons que deux vétérinaires évaluent le degré de sédation de 56 chats après prémédication avec de la méthadone. Supposons maintenant deux situations n°1 et n°2. Les tableaux 7.1 et 7.2 présentent la répartition des degrés de sédation dans chacune de ces deux situations. Ces tableaux montrent que la concordance entre vétérinaires sur le degré de sédation est plus grande dans la situation n°1 que dans la situation n°2. En effet, les seules différences entre les deux tableaux, exprimées en gras et soulignées, sont les suivantes. Dans la situation n°1, 8 sédations sont jugées « modérées » par le vétérinaire 1 et « importantes » par le vétérinaire 2 (ce qui n'est pas *très* discordant), et 1 seule sédation est jugée « légère » par le vétérinaire 1 et « importante » par le vétérinaire 2 (ce qui est très discordant) ; dans la situation n°2, 8 sédations sont jugées « légères » par le vétérinaire 1 et « importantes » par le vétérinaire 2 (ce qui est très discordant), et 1 sédation est jugée « modérée » par le vétérinaire 1 et « importante » par le vétérinaire 2 (ce qui n'est pas *très* discordant). Par conséquent, d'un point de vue clinique, la situation n°2 présente beaucoup plus de degrés de sédation *très* discordants que la situation n°1 (8 *versus* 1, respectivement). Mais dans la mesure où le caractère est considéré comme nominal avec le calcul du coefficient de concordance Kappa classique, il n'y a pas de gradation dans la discordance ; ces deux situations sont de la même façon non concordantes, et fournissent par conséquent des valeurs quasi identiques du coefficient Kappa (respectivement de +0,53 et +0,51, pour les situations n°1 et n°2).

Tableau 7.1. Degré de sédation évalué par deux vétérinaires sur 56 chats (situation n°1).

		Degré de sédation, vét. n°1				
		Absente	Légère	Modérée	Importante	Total
Degré de sédation, vét. n°2	Absente	6	2	0	0	8
	Légère	1	14	4	0	19
	Modérée	0	1	2	2	5
	Importante	0	**1**	**8**	15	21
	Total	7	18	14	17	56

Tableau 7.2. Degré de sédation évalué par deux vétérinaires sur 56 chats (situation n°2).

		Degré de sédation, vét. n°1				
		Absente	Légère	Modérée	Importante	Total
Degré de sédation, vét. n°2	Absente	6	2	0	0	8
	Légère	1	14	4	0	19
	Modérée	0	1	2	2	5
	Importante	0	**8**	**1**	15	24
	Total	7	25	7	17	56

Une solution face à ce « paradoxe » (qui n'en est pas un si l'on a en tête que le coefficient de concordance Kappa classique considère le caractère qualitatif comme un caractère qualitatif *nominal*) est de regrouper, par exemple, les sédations « absentes » et « légères » ensemble, puis les « modérées » et « importantes » ensemble.

Les tableaux 8.1 et 8.2 fournissent les nouvelles répartitions, pour respectivement les situations n°1 et n°2.

Tableau 8.1. Caractère qualitatif ordinal du tableau 7.1 rendu binaire.

		Degré de sédation, vét. 1		
		Absente, légère	Modérée, importante	Total
Degré de sédation, vét. 2	Absente, légère	23	4	27
	Modérée, importante	2	27	29
	Total	25	31	56

Tableau 8.2. Caractère qualitatif ordinal du tableau 7.2 rendu binaire.

		Degré de sédation, vét. 1		
		Absente, légère	Modérée, importante	Total
Degré de sédation, vét. 2	Absente, légère	23	4	27
	Modérée, importante	9	20	29
	Total	32	24	56

Dans la situation n°1 en regroupant les degrés de sédation (tableau 8.1), la valeur du coefficient de concordance Kappa est de +0,78. Dans la situation n°2 en regroupant les degrés de sédation (tableau 8.2), la valeur du coefficient de concordance Kappa est de +0,54, valeur bien inférieure à celle de la situation n°1, ce qui redevient cohérent avec les observations cliniques (plus grande concordance clinique entre vétérinaires dans la situation n°1 que dans la situation n°2).

Une autre solution consiste à calculer un coefficient Kappa « pondéré », qui permet de prendre en compte l'ordre des classes dans les calculs (Cohen, 1968; Sim and Wright, 2005). La pondération va pénaliser le coefficient de concordance Kappa (en le rapprochant de 0) d'autant plus que les discordances sont importantes (éléments discordants loin de la diagonale ; cf. valeur de « 8 » dans le tableau 7.2). Deux types de pondérations sont les plus fréquemment utilisés : la pondération linéaire et la pondération quadratique. La pondération consiste à affecter un poids aux éléments hors de la diagonale (éléments discordants) d'autant plus importante (linéairement ou quadratiquement, en fonction du type de pondération) que l'on s'éloigne de cette diagonale. La pondération quadratique pénalise moins le coefficient de concordance Kappa pour de petits écarts de concordance (relativement au nombre total de classes) que la pondération linéaire ; elle pénalise plus le coefficient de concordance Kappa pour de grands écarts de concordance (relativement au nombre total de classes) que la pondération linéaire. Tout dépend donc si vous souhaitez peu pénaliser les petits écarts et beaucoup les grands écarts ($\rightarrow$ choix de la pondération quadratique) ou pénaliser de façon homogène les écarts ($\rightarrow$ pondération linéaire). Cependant, compte tenu du fait que la valeur du coefficient de concordance Kappa pondéré dépend du nombre de catégories davantage avec une pondération quadratique qu'avec une pondération linéaire (Brenner and Kliebsch, 1996), la pondération linéaire devrait être préférentiellement utilisée, surtout si le nombre de catégories est important. Mais la littérature n'est quand même pas claire sur ce dernier point (Sim and Wright, 2005) ! Enfin, il est à noter que des limites dans l'utilisation du coefficient Kappa pondéré ont été décrites dans l'article et Graham et Jackson de 1993 (Graham and Jackson, 1993).

En reprenant les tableaux 7.1 et 7.2, les valeurs du coefficient Kappa pondéré linéairement sont +0,70 et +0,60 respectivement pour les situations n°1 et n°2 ; elles sont de +0,83 et +0,68 en utilisant la pondération quadratique respectivement pour les situations n°1 et n°2. Ainsi, que ce soit en utilisant la pondération linéaire ou la pondération quadratique, les résultats sont cohérents avec les observations cliniques (coefficients de concordance Kappa supérieurs pour la situation n°1 par rapport à la situation n°2).

IV. Concordance entre deux séries de mesures quantitatives

A. Introduction

Il existe des méthodes numériques et des méthodes graphiques pour évaluer la répétabilité/reproductibilité d'une méthode de mesure, ou la concordance entre deux méthodes de mesure, lorsque le caractère est quantitatif. La méthode **numérique** présentée dans ce guide pratique est le coefficient de concordance de Lin. Les méthodes graphiques quant à elles représentent graphiquement la concordance ; elles permettent d'interpréter *cliniquement* la concordance de deux séries de mesures. Ces dernières peuvent ainsi nuancer, infirmer, ou bien au contraire confirmer le niveau de concordance quantifié par les méthodes numériques. La méthode **graphique** présentée dans ce guide pratique est la méthode de Bland et Altman.

B. Quantification numérique de la concordance de deux séries de mesures

Les méthodes numériques que l'on rencontre dans la littérature permettant *a priori* de quantifier la concordance entre deux séries de mesures quantitatives sont les suivantes : le coefficient de variation, les coefficients de corrélation de Pearson et de Spearman, les comparaisons de séries appariées avec le test de Student, l'analyse des moindres carrés en traçant une droite d'équation y = a.x + b (en testant a et b), le coefficient de corrélation intraclasse, et le coefficient de concordance de Lin. Nous allons voir que certaines de ces méthodes ne permettent pas rigoureusement de quantifier la concordance entre deux séries de mesures quantitatives.

1. Le coefficient de variation

Le coefficient de variation est un indicateur qui quantifie la part de la variabilité d'un caractère quantitatif mesuré plusieurs fois rapportée à la moyenne de ce caractère calculée à partir de ces mêmes mesures :

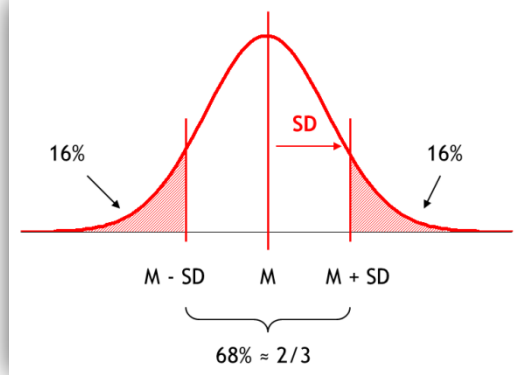
$$CV = \frac{Standard\ Deviation}{Mean}$$

Certains chercheurs tiennent le raisonnement suivant : si, après avoir mesuré *N* fois un individu, le CV calculé sur ces *N* mesures est faible (< 10%), alors la méthode de mesure peut être considérée comme répétable. Cette interprétation du CV doit être cependant prise avec beaucoup de précautions.

Voici tout d'abord quelques rappels de statistique de base. Supposons un caractère quantitatif suivant parfaitement une loi normale, centrée sur la moyenne M calculée dans l'échantillon, et d'écart-type[2] SD calculé lui aussi dans l'échantillon. Alors, on peut dire qu'environ 2/3 des individus de l'échantillon[3] ont une valeur de ce caractère comprise entre M-SD et M+SD (cf. figure ci-dessous).

[2] « Standard Deviation » en anglais

[3] 68% pour être exact



Par conséquent, si la valeur du CV = 10%, cela signifie qu'environ 2/3 des mesures ont une valeur comprise entre M-0,10xM et M+0,10xM, soit dans l'intervalle {M ± 10% de M}. Cela veut donc aussi dire qu'environ 1/3 des mesures ont une valeur à l'*extérieur* de cet intervalle. Peut-on considérer que ce « 1/3 » est une faible proportion ?... Non ! Par conséquent, une valeur de CV < 10% n'indique pas forcément que les *N* valeurs mesurées sur un individu sont proches les unes des autres.

Deuxième problème. On ne travaille que sur un seul individu (dont la moyenne des valeurs de chaque mesure effectuée est M). Mais quelle est la représentativité de cet individu vis-à-vis des individus de la population cible ? C'est difficile à dire. La valeur du CV est donc propre à l'individu. Un individu a sa propre moyenne et son propre écart-type. Est-il possible de généraliser cette valeur de CV à l'ensemble des individus, pour pouvoir dire que la méthode de mesure est répétable ?... C'est donc probablement difficile à faire.

Enfin, la valeur de CV dépend de M. Pour une autre valeur de M, à écart-type constant, la valeur de CV sera différente. Dans ces conditions, il est difficile de généraliser en disant que la méthode est répétable avec une faible valeur de CV. Plus d'informations sur les limites de l'interprétation du CV peuvent être trouvées dans l'article de Lachin (Lachin, 2004).

2. *Les coefficients de corrélation*

Un coefficient de corrélation quantifie l'association entre deux variables quantitatives. Si ces deux variables quantitatives suivent une loi normale, alors le coefficient de corrélation de Pearson peut être utilisé ; si ce n'est pas le cas, le coefficient de corrélation de Spearman peut être utilisé. Le coefficient de corrélation varie entre -1 et +1.

La valeur absolue d'un coefficient de corrélation égale à 1 signifie qu'il existe une association parfaite entre les deux variables (soient X et Y ces deux variables). Qu'est-ce qu'une association entre X et Y parfaite? C'est lorsque les points d'abscisse X et d'ordonnée Y sont tous sur une même droite, quelle que soit la pente de la droite (à part la pente nulle). La valeur absolue d'un coefficient de corrélation est d'autant plus loin de 1 et proche de 0 que les points sont éloignés de la droite de régression passant par ces points (celle dressée de telle façon à ce que les écarts entre les points et la droite soient les plus faibles possibles).

Certains chercheurs tiennent le raisonnement suivant : si le coefficient de corrélation (de Spearman ou de Pearson) quantifiant l'association entre deux séries de mesures est très proche de +1, alors ces deux séries de mesures sont très concordantes. Pour illustrer la raison pour laquelle ce raisonnement est faux (une valeur d'un coefficient de corrélation très proche de +1 ne signifie pas que les deux séries de mesures sont très concordantes), je vais utiliser l'exemple ci-dessous.

Supposons que l'on ait 5 méthodes de mesure pour mesurer la concentration en créatinine C_{creat} chez le chien : une méthode de référence (M_REF) et quatre méthodes (M1, M2, M3, et M4) dont on voudrait savoir si elles donnent des résultats concordants avec la méthode de référence (cf. tableau 9).

La figure 2 représente les mesures de C_{creat} effectuées sur les 15 prélèvements sanguins, avec en abscisse la valeur de C_{creat} mesurée par la méthode de référence, et en ordonnée la valeur de C_{creat} pour les mêmes prélèvements mesurée par les méthodes M1 et M2. La droite à 45° représente le fait que si les méthodes M1 et M2 donnaient des valeurs de C_{creat} identiques à la méthode de référence, alors les cercles pleins (M1) et les cercles vides (M2) devraient se trouver sur cette droite. Cette droite à 45° représente donc la concordance parfaite avec la méthode de référence M_REF.

Tableau 9. Valeurs de la concentration plasmatique en créatinine (en mg/dl) mesurées par 5 méthodes de mesure (la méthode de référence M_REF, et les quatre méthodes M1, M2, M3, et M4) à partir de 15 prélèvements sanguins de chiens.

N° du chien	M_REF	M1	M2	M3	M4
1	0,50	0,81	0,47	0,61	0,56
2	0,55	0,81	0,62	0,73	0,61
3	0,60	0,92	0,75	0,81	0,35
4	0,65	0,99	0,84	0,77	0,84
5	0,70	0,99	0,98	0,76	0,52
6	0,75	1,09	1,02	0,95	0,97
7	0,80	1,09	1,19	0,97	0,83
8	0,85	1,12	1,24	0,97	1,13
9	0,90	1,23	1,36	1,01	0,86
10	0,95	1,23	1,48	1,14	1,08
11	1,00	1,27	1,57	1,2	0,76
12	1,05	1,35	1,71	1,33	1,06
13	1,10	1,39	1,77	1,29	1,35
14	1,15	1,43	1,95	1,26	1,14
15	1,20	1,49	2,02	1,4	1,22

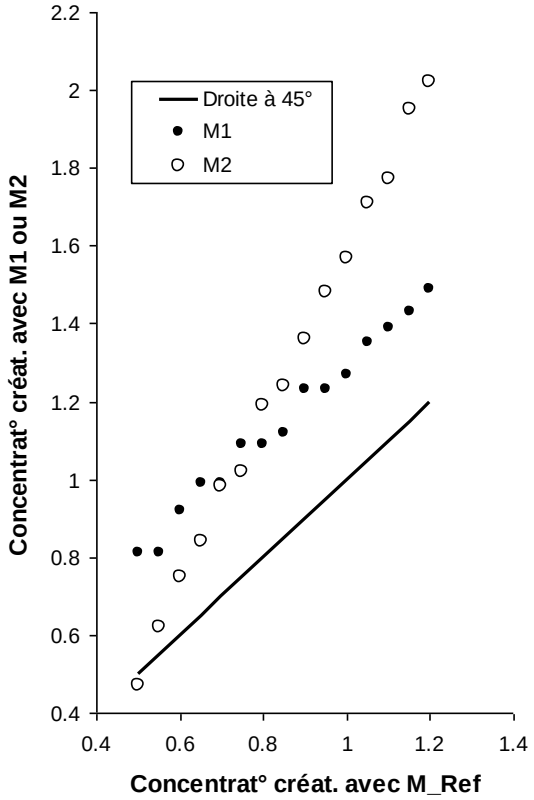


Figure 2. Représentation graphique des concentrations plasmatiques en créatinine (C_{creat}, en mg/dl) mesurées sur 15 prélèvements sanguins de chiens, avec en abscisse la valeur de C_{creat} mesurée par la méthode de référence. En ordonnée se trouve la valeur de C_{creat} mesurée par les méthodes M1 et M2. La droite en trait plein est celle à 45° ; elle représente la concordance parfaite avec la méthode de référence.

A l'aide de la figure 2, on se rend aisément compte que les deux méthodes M1 et M2 ne fournissent pas des résultats concordants avec la méthode de référence (les points sont loin de la droite de concordance parfaite à 45°). Pourtant, si l'on calcule le coefficient de corrélation de Pearson quantifiant la corrélation entre les mesures effectuées avec la méthode de référence et celles avec M1, on obtient une valeur de 0,99 ; le coefficient de corrélation de Pearson est égal à 1,00 pour la corrélation entre les mesures effectuées avec M2 et celles effectuées avec la méthode de référence[4]. Par conséquent, si l'on se limitait au calcul du coefficient de corrélation de Pearson, on conclurait totalement à tort que, dans la mesure où la *corrélation* entre mesures est (presque) parfaite, les méthodes M1 et M_REF ou les méthodes M2 et M_REF sont totalement équivalentes, et donc parfaitement interchangeables ! (Ce qui est bien entendu faux au vu de la figure 2.)

Pour information, la raisonnement qui consiste à tester le coefficient b dans une droite de régression linéaire dont la formule serait M_Ref = a + b.M_1, puis à dire que, puisque b est significativement différent de 0, cela signifie que la méthode de référence est concordante à la méthode M1, est basé sur la même erreur de raisonnement que celui sur le coefficient de corrélation.

En conclusion, ni les coefficients de corrélation, ni les droites de régression linéaire ne fournissent d'information pertinente pour évaluer la concordance entre deux séries de mesures.

[4] Rappelons que la valeur maximale pour un coefficient de corrélation est égale à 1,00, indiquant une corrélation parfaite entre les deux variables quantitatives.

3. *La comparaison statistique des deux séries appariées à l'aide d'un test de Student*

Certains chercheurs (pas forcément les mêmes que les précédents ☺) tiennent le raisonnement suivant : si la moyenne de la concentration en créatinine mesurée avec la méthode de référence n'est pas significativement différente de la moyenne de la concentration en créatinine mesurée avec la méthode M1 (ou M2)[5], alors les deux méthodes peuvent être considérées comme concordantes. Ce raisonnement est faux pour deux raisons. La première, l'acceptation de l'hypothèse nulle dans un test statistique (lorsque $p > 0,05$) est assortie d'un risque d'erreur de 2ème espèce β dont la valeur est inconnue (puisqu'elle dépend de la vraie différence de moyennes entre les deux populations comparées). Ainsi, il n'est pas question d'accepter *avec force* l'hypothèse nulle (dire avec force que « les deux moyennes ne sont pas significativement différentes, donc elles sont égales dans la population »). Deuxième raison, les deux moyennes peuvent être égales avec deux séries complètement discordantes. L'exemple ci-dessous le prouve : les trois individus mesurés n'ont pas la même valeur d'une série à l'autre, et pourtant, la moyenne de la série n°1 est égale à celle de la série n°2 (toutes deux égales à 30).

N° du chien	Série n°1	Série n°2
1	20	40
2	40	10
3	30	40

4. *Le coefficient de concordance de Lin*

Nous venons donc de voir que ni le coefficient de corrélation, ni le coefficient directeur d'une droite de régression, ni la comparaison statistique des deux moyennes des séries ne permettent de correctement évaluer la concordance de deux séries de mesures. L'une des méthodes numériques préconisées est celle du calcul du coefficient de concordance de Lin (Barnhart et al., 2002; Lin, 1989).

Cela dit, le coefficient de concordance de Lin ne peut être utilisé que dans la situation où, en toute théorie et dans un monde idéal, les deux séries de mesures sont censées être identiques (concordance parfaite). Ainsi, la situation où les deux séries de mesures quantitatives ne sont pas exprimées dans la même unité et/ou ne sont *a priori* pas censées fournir les mêmes valeurs (par exemple, la concentration en cortisol en nmol/l, entre la concentration salivaire et la concentration plasmatique), le coefficient de concordance de Lin ne doit pas être calculé. En effet, dans cette situation, la question sera de savoir si les deux séries sont *corrélées* (et non pas *concordantes*). Il faudra alors calculer un coefficient de corrélation (de Pearson ou de Spearman).

Le coefficient de concordance de Lin est un coefficient allant de -1 à +1, où les valeurs de -1, 0, et +1 signifient respectivement une discordance parfaite, une concordance nulle, et une concordance parfaite. De façon simple, le coefficient de concordance de Lin quantifie les écarts entre les points d'abscisse la série n°1 et d'ordonnée la série n°2 et la droite à 45° représentant la concordance parfaite (cf. figure 2).

[5] En utilisant le test de Student pour séries appariées.

La formule du coefficient de concordance de Lin est la suivante (Lin, 1989) :

$$\rho_c = \frac{2\sigma_{12}}{\sigma_1^2 + \sigma_2^2 + (\mu_1 - \mu_2)^2} = \rho C_b,$$

where

$$C_b = [(v + 1/v + u^2)/2]^{-1},$$

$v = \sigma_1/\sigma_2$ = scale shift,

$u = (\mu_1 - \mu_2)/\sqrt{\sigma_1 \sigma_2}$ = location shift relative to the scale.

Avec :

ρ_C = coefficient de concordance de Lin

ρ = coefficient de corrélation de Pearson

C_b = un coefficient dépendant de l'écart de variabilité du caractère quantitatif entre les deux séries, et de l'écart de moyennes entre les deux séries. Ce coefficient C_b est compris entre 0 et +1. Il quantifie (de façon inverse) l'écart entre la droite de régression estimée à partir des deux séries de mesures et la droite à 45°. Ce coefficient vaut « 1 » si la droite de régression est identique à la droite à 45° ; plus l'écart entre la droite de régression et la droite à 45° augmente, plus C_b se rapproche de « 0 » (Lin, 1989).

Cette formule nous permet déjà de faire la remarque suivante : le coefficient de concordance de Lin ne peut pas être supérieur au coefficient de corrélation de Pearson (puisque $C_b \leq 1$).

Si l'on reprend la figure 2, on peut voir que les points sont éloignés de la droite à 45°. Ces écarts à la droite à 45° se traduisent par des valeurs du coefficient de concordance de Lin bien inférieures à 1, traduisant une mauvaise concordance entre la méthode de référence et les méthodes M1 ou M2 : respectivement 0,52 (contre 0,99 pour le coefficient de corrélation de Pearson) et 0,47 (contre 1,00 pour le coefficient de corrélation de Pearson).

Introduisons désormais les notions de *précision* et d'*exactitude* de la concordance. Le coefficient de corrélation de Pearson quantifie la relation linéaire qui existe entre les deux séries de mesures. Si les mesures observées s'éloignent de part et d'autre de la droite de régression estimée à partir des deux séries de mesures, on va dire qu'il y a un manque de « précision » autour de cette droite de régression (on parle de manque de *précision* de la concordance, et ce manque est quantifié, de façon inverse, par le coefficient de corrélation de Pearson) ; dans ce cas, le coefficient de corrélation de Pearson se rapproche de « 0 ». Maintenant, si la droite de régression estimée à partir des deux séries de mesures est éloignée de la droite à 45° représentant la concordance parfaite, on va dire qu'il y a un manque d' « exactitude » autour de la droite à 45° (on parle de manque d'*exactitude* de la concordance, et ce manque est quantifié, de façon inverse, par le coefficient C_b) ; dans ce cas, le coefficient C_b se rapproche de « 0 ». Ainsi, la valeur du coefficient de concordance de Lin est pénalisée tout autant par un manque de *précision* de la concordance que par un manque d'*exactitude* de la concordance (Lin, 2000).

Il peut être intéressant de calculer la part du manque de *précision* et celle du manque d'*exactitude* de la concordance, quand on observe un coefficient de concordance de Lin éloigné de « 1 ». Les formules ci-dessous[6] permettent de calculer la part (en %) du manque de précision et celle du manque d'exactitude dans le fait que le coefficient de concordance de Lin est éloigné de « 1 ».

[6] Ces formules ne sont pas issues de référence bibliographiques, mais sont issues de ma réflexion seulement. L'idée a été la suivante : quels sont les « poids » que l'on peut attribuer à A et à B dans le résultat de C, si C = A x B avec A et B $\in$]0 ; 1], et

Part (en %) du manque de précision = $\frac{Ln(\rho)}{Ln(\rho)+Ln(C_b)}$

Part (en %) du manque d'exactitude = $\frac{Ln(C_b)}{Ln(\rho)+Ln(C_b)}$

Les figures 3.a et 3.b ci-dessous sont une copie d'écran des informations que fournit le fichier Excel® calculant le coefficient de corrélation de Pearson, le coefficient C_b, et le coefficient de concordance de Lin assorti de son intervalle de confiance à 95%, pour quantifier la concordance entre la série de mesures avec la méthode de référence M_REF (S1, colonne A) et la série de mesures avec la méthode M3 (S2, colonne B dans la figure 3.a) ou la série de mesures avec la méthode M4 (S2, colonne B dans la figure 3.b). Les valeurs de ces séries de mesures avec les méthodes M3 et M4 se retrouvent aussi dans le tableau 9.

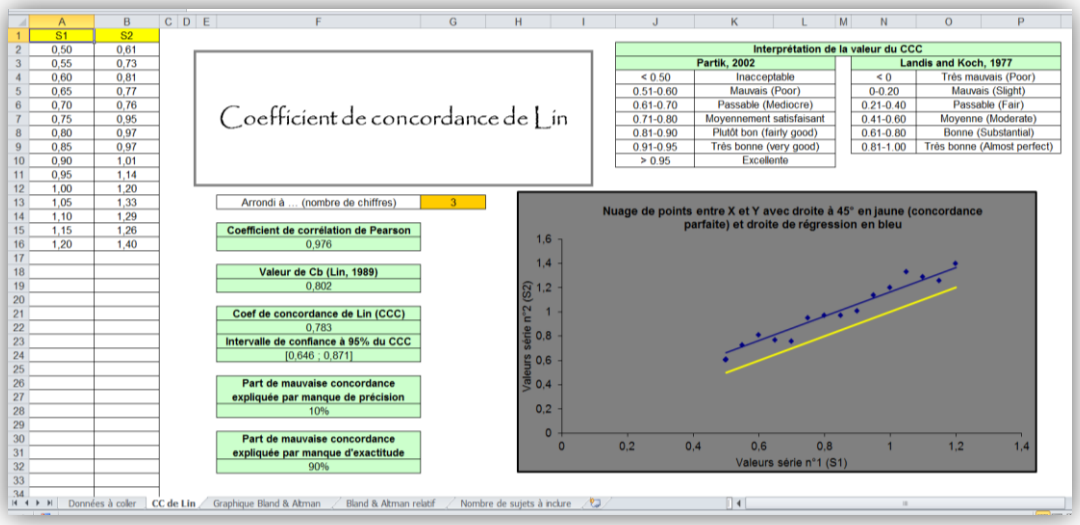


Figure 3.a. Valeurs du coefficient de corrélation de Pearson, de celle de C_b, et de celle du coefficient de concordance de Lin (avec son intervalle de confiance à 95%) correspondant aux concentrations de la créatinine de 15 chiens évaluées par les méthodes M_REF et M3 (cf. tableau 9).

Sur la figure 3.a, on peut lire la valeur du coefficient de corrélation de Pearson (0,976), celle du coefficient C_b (0,802), et le coefficient de concordance de Lin (0,783 = 0,976x0,802). La droite bleue représente la droite de régression passant par les points, et la droite jaune représente la droite à 45° (les points bleus auraient dû être sur la droite jaune s'il y avait une concordance parfaite entre la méthode M_REF et la méthode M3).

de telle sorte qu'une valeur de A (ou B) proche de 0 doit augmenter ce « poids ». Après différents essais, j'ai sélectionné les formules utilisant le logarithme népérien. (L'échelle logarithmique est plus naturelle car elle transforme un produit en une addition : Ln(C) = Ln(A) + Ln(B)).

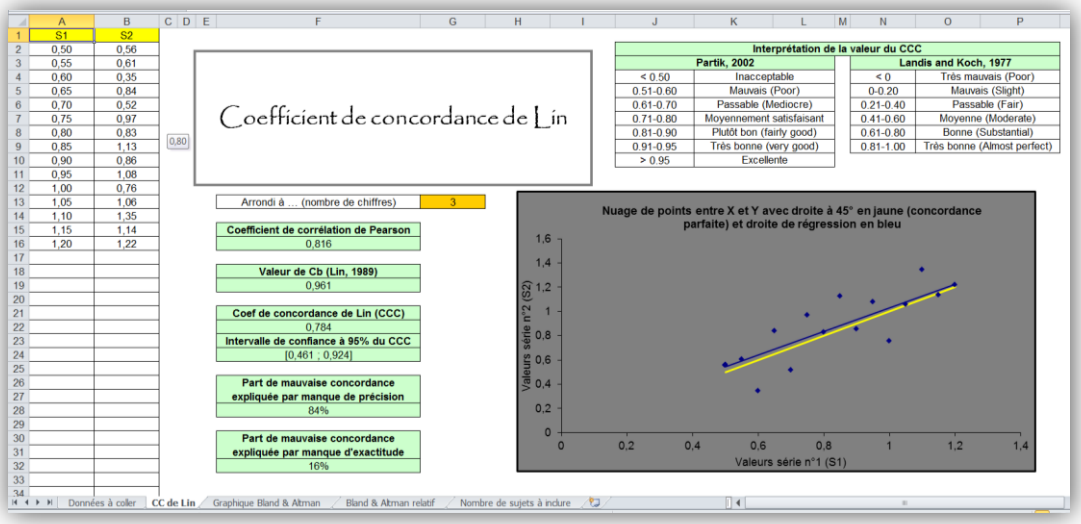


Figure 3.b. Valeurs du coefficient de corrélation de Pearson, de celle de C_b, et de celle du coefficient de concordance de Lin (avec son intervalle de confiance à 95%) correspondant aux concentrations de la créatinine de 15 chiens évaluées par les méthodes M_REF et M4 (cf. tableau 9).

Sur la figure 3.b, on peut lire la valeur du coefficient de corrélation de Pearson (0,816), celle du coefficient C_b (0,961), et le coefficient de concordance de Lin (0,784 = 0,816x0,961).

Cherchons maintenant à identifier la cause d'un coefficient de concordance de Lin éloigné de « 1 », que ce soit dans l'évaluation de la concordance des méthodes M_REF et M3 ou M_REF et M4 : s'agit-il plutôt d'un manque de précision de la concordance, d'un manque d'exactitude de la concordance, on bien d'un manque des deux ?

Concernant la concordance entre les méthodes M_REF et M3, la part du manque de précision est de Ln(0,976)/(Ln(0,976)+Ln(0,802)) ≈ 10%, et la part du manque d'exactitude est de Ln(0,802)/(Ln(0,976)+Ln(0,802)) ≈ 90%[7]. Ces deux valeurs sont présentées dans la feuille Excel®, sous les valeurs de l'intervalle de confiance du coefficient de concordance de Lin. Ainsi, le manque de concordance entre la méthode de référence et la méthode M3 provient bien davantage d'un manque d'exactitude (90%) que d'un manque de précision (10%). La droite de régression (en bleu) étant au dessus de la droite à 45° (en jaune), cela indique que ce manque d'exactitude est une sur-estimation de la méthode M3 par rapport à la méthode de référence sur toute la plage des valeurs, qu'il faudrait donc corriger.

Concernant la concordance entre les méthodes M_REF et M4, la part du manque de précision est de Ln(0,816)/(Ln(0,816)+Ln(0,961)) ≈ 84%, et la part du manque d'exactitude est de Ln(0,961)/(Ln(0,816)+Ln(0,961)) ≈ 16%. Ainsi, le manque de concordance entre la méthode de référence et la méthode M4 provient bien davantage d'un manque de précision (84%) que d'un manque d'exactitude (16%). Pour améliorer la concordance entre la méthode de référence et la méthode M4, il faudrait réduire les erreurs aléatoires, c'est-à-dire rendre plus précis le protocole de mesure de la méthode M4.

En résumé, la valeur de ces deux parts (part du manque de précision et part du manque d'exactitude) permet de prioriser les actions pour améliorer la concordance entre les deux séries de mesures : améliorer la précision (réduire les erreurs aléatoires entre les deux séries) et/ou améliorer l'exactitude (réduire les erreurs systématiques entre les deux séries de mesures). Pour savoir comment réduire les erreurs systématiques, il est indispensable de comparer la droite de régression

[7] En utilisant les deux formules précédentes.

à la droite à 45°. C'est cet écart entre les deux droites qui vous guidera dans les mesures correctives à apporter.

L'interprétation de la valeur du coefficient de concordance de Lin est identique à celle du coefficient de concordance Kappa. Ainsi, nous pouvons utiliser la classification subjective de Landis et Koch pour dire si la concordance est moyenne, bonne, ou excellente (cf. tableau 4) (Donner and Eliasziw, 1987). Partik et coll. ont proposé une autre classification, qui est moins souvent reprise dans la littérature, mais que je vous présente dans le tableau 10 ci-dessous (Partik et al., 2002).

Tableau 10. Interprétation des valeurs du coefficient de concordance de Lin (CC Lin) (Partik et al., 2002).

CC Lin	Interprétation
< 0,50	Inacceptable
0,51-0,60	Poor
0,61-0,70	Mediocre
0,71-0,80	Satisfactory
0,81-0,90	Fairly good
0,91-0,95	Very good
> 0,95	Excellent

5. *Le coefficient de corrélation intraclasse*

Le coefficient de corrélation intraclasse est un autre indicateur pour quantifier la concordance de séries de mesures quantitatives (Lee et al., 1989; Muller and Buttner, 1994; Shrout and Fleiss, 1979). Cependant, j'ai choisi de ne pas présenter cet indicateur car (1) de nombreux ICC existent et le choix de l'un vis-à-vis des autres peut être difficile, (2) ils nécessitent des calculs statistiques plus compliqués que ceux du coefficient de concordance de Lin (qui lui est calculable à partir d'un fichier Excel®), et (3) les valeurs de ces deux coefficients sont très proches (Carrasco and Tover, 2003; Chen and Barnhart, 2008; Nickerson, 1997).

6. *« Reliability » versus « agreement »*

Les articles suivants présentent en détails la distinction entre « reliability » et « agreement » (Bartlett and Frost, 2008; de Vet et al., 2006; Kottner and Streiner, 2011). Une méthode de mesure a une bonne « reliability » si elle est capable de correctement distinguer deux individus, malgré l'erreur de mesure (erreur systématique[8] + erreur aléatoire). Une méthode de mesure a un bon « agreement » si elle est capable de fournir deux fois la même valeur chez un individu mesuré deux fois, indépendamment des valeurs des individus.

Le coefficient de concordance de Lin ainsi que le coefficient de corrélation intraclasse quantifient tous les deux davantage la « reliability » plutôt que l' « agreement » (de Vet et al., 2006). En effet, ces deux coefficients sont proportionnels au ratio hétérogénéité individuelle / erreur de mesure. Plus ce ratio est élevé, moins le bruit dû à l'erreur de mesure perturbe la distinction entre deux individus, et donc plus les coefficients de concordance de Lin et de corrélation intraclasse seront proches de 1. La figure 4 illustre ce que je viens d'écrire.

[8] L'erreur systématique est celle qui correspond à une sur-estimation (ou sous-estimation) systématique de la série n°2 par rapport à la série n°1 (notamment dans le cas de reproductibilité ou de concordance entre deux méthodes de mesures).

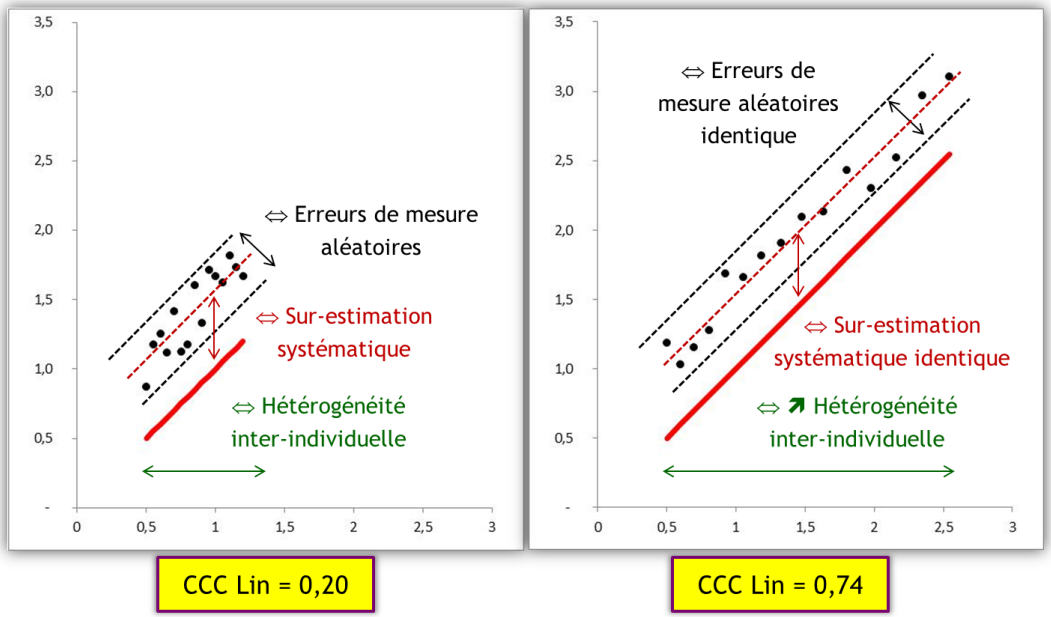


Figure 4. Illustration du fait que le coefficient de concordance de Lin quantifie davantage la « reliability » que l' « agreement ». Dans les deux situations, l'erreur systématique est la même (la série 2 sur-estime la série 1 de la même quantité en moyenne), avec des erreurs aléatoires comparables, mais puisque l'hétérogénéité inter-individuelle est plus élevée sur la figure de droite que sur la figure de gauche, le coefficient de concordance de Lin est plus élevé. Cette situation serait la même avec le calcul du coefficient de corrélation intraclasse.

Par conséquent, quand la valeur du coefficient de concordance de Lin est élevée, il n'y a pas trop de questions à se poser, les deux séries sont concordantes. En revanche, quand ce coefficient commence à avoir des valeurs considérées comme non satisfaisantes (cf. tableau 4), il devient difficile de déterminer la cause de ces faibles valeurs (Atkinson and Nevill, 1997) : l'erreur de mesure aléatoire est-elle trop importante, l'erreur systématique est-elle trop importante, ou bien existe-t-il une variabilité inter-individuelle trop faible au regard de l'erreur de mesure ?

Cette considération ne doit pas empêcher le calcul du coefficient de concordance de Lin dans la validation de la méthode de mesure, mais le point évoqué ci-dessus doit être gardé en mémoire au moment de l'interprétation de faibles valeurs du coefficient de concordance de Lin.

La solution que préconisent les différents auteurs ayant relevé cette distinction entre « reliability » et « agreement » pour vraiment quantifier l'agreement est d'utiliser la méthode graphique de Bland et Altman.

C. Appréciation graphique de la concordance de deux séries de mesures à l'aide de la méthode de Bland et Altman

1. *Introduction*

Comme nous venons de le voir dans l'illustration de partie précédente, deux valeurs de coefficient de concordance de Lin similaires peuvent traduire des situations de non concordance assez différentes : la méthode de mesure M1 sur-estimait systématiquement la C_{creat} de 0,3 mg/dl, tandis que la méthode de mesure M2 sur-estimait la C_{creat} de façon proportionnelle à la valeur réelle de C_{creat} (figure 2), bien que conduisant à des valeurs de coefficient de concordance de Lin assez proches (0,52 et 0,47, respectivement). C'est en effet la limite des indicateurs numériques (une même valeur d'un indicateur n'est pas forcément le reflet d'une même situation clinique). La méthode graphique de Bland et Altman est une méthode « clinique » d'évaluation de la concordance entre deux séries de

mesures, car les critères de concordance sont des critères « cliniques », fixés *a priori* par l'investigateur (alors que le coefficient de concordance de Lin est un critère « statistique » – cf. tableaux 4 et 10). Un excellent exemple d'utilisation de la méthode de Bland et Altman pour évaluer *cliniquement* la répétabilité d'une méthode de mesure se trouve dans l'article de Bakker et al. (Bakker et al., 1999). D'autres méthodes graphiques évaluant la concordance entre deux séries de mesures existent (Krouwer and Monti, 1995; Luiz et al., 2003), mais ne sont pas autant utilisées que celle de Bland et Altman.

Par ailleurs, et de la même façon que pour le coefficient de concordance de Lin, en toute théorie et dans un monde idéal, les deux séries de mesures sont censées être identiques (concordance parfaite).

Pour présenter la méthode de Bland et Altman, je vais fortement m'inspirer de la thèse d'exercice vétérinaire du Dr C Boyer, dirigée par le Dr M Huynh, en utilisant les données recueillies pour la thèse (modifiées pour des raisons pédagogiques).

2. *Présentation de la méthode graphique de Bland et Altman*

La méthode graphique de Bland et Altman repose sur la définition même de la concordance entre deux séries de mesures (Bland and Altman, 1986). Les deux séries sont concordantes si l'une ne sur-estime ou ne sous-estime pas l'autre de façon trop importante, *et* si les écarts entre les deux séries pour chaque individu mesuré (deux fois) ne sont pas trop importants (cf. figure 5). Vous vous rendez compte d'emblée le caractère subjectif de la méthode, avec l'expression « pas trop important(e) ». Ca va être en effet à l'investigateur de l'étude de fixer *a priori* ce qu'il juge comme « pas trop important ».

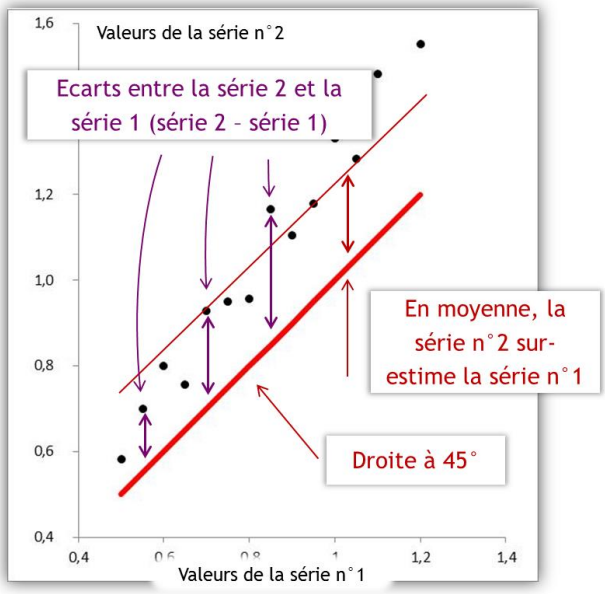


Figure 5. Illustration des deux critères de concordance dans la méthode de Bland et Altman : l'écart moyen entre la série de mesures n°2 et la série de mesures n°1 (écart entre la droite rouge en trait fin et la droite à 45°), et l'ensemble des écarts entre les deux séries de mesures.

Le graphique de Bland et Altman est composé de trois éléments : des points représentant les mesures des individus (il y a autant de points que d'individus mesurés deux fois), une droite centrale, et deux droites « extérieures » (Bland and Altman, 1999). La figure 6 ci-dessous est issue de l'article de Voyvoda et coll. (Voyvoda and Erdogan, 2010), dont l'objectif était de quantifier la concordance de deux méthodes de mesure de la concentration en corps cétoniques (BHBA) chez la vache : la méthode « Laboratory » (série n°1) et la méthode « Hand-Held meter » (série n°2).

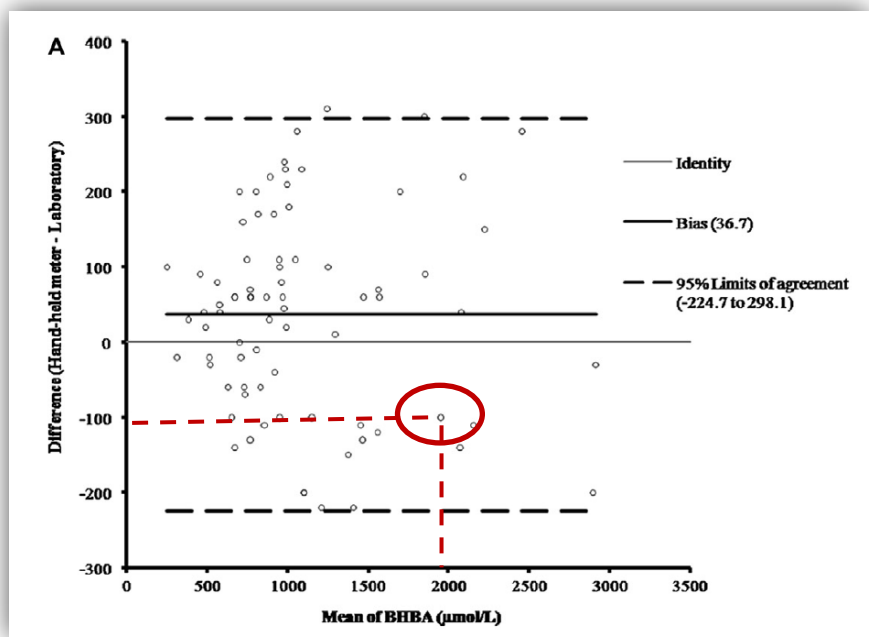
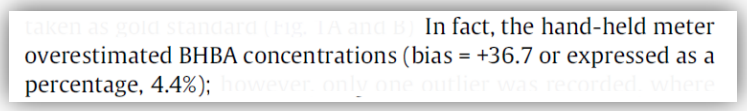


Figure 6. Graphique de Bland et Altman issue de l'article de Voyvoda et coll. (Voyvoda and Erdogan, 2010). Le point entouré correspond à une vache mesurée deux fois (voir texte pour plus d'explications).

Dans le graphique de Bland et Altman, l'axe des abscisses est la moyenne des mesures des deux séries, et l'axe des ordonnées est la différence entre les mesures des deux séries (la série n°2 – la série n°1). Le point entouré sur la figure 6 a pour abscisse ≈ 1968 µmol/L et pour ordonnée ≈ -106 µmol/L. Cela signifie que, pour cette vache-là, la moyenne de valeurs de BHBA mesurée par la méthode « Laboratory » et par la méthode « Hand-Held meter » est égale à 1968 µmol/L, et que la différence entre les deux valeurs est de -106 µmol/L. On en déduit les deux valeurs pour cette vache-là[9] : la valeur de la concentration en BHBA avec la méthode « Laboratory » était de 2021 µmol/L, et celle avec la méthode « Hand-Held meter » était de 1915 µmol/L. Ainsi, on peut se rendre compte que plus le point est éloigné de l'axe des abscisses (appelé « Identity » sur la figure 6), plus l'écart entre la série n°1 et la série n°2 pour l'individu correspondant est important, que ce soit une sur-estimation de la série n°2 par rapport à la série n°1 (point au dessus de l'axe des abscisses) ou une sous-estimation (point au dessous de l'axe des abscisses).

[9] En résolvant les deux équations à deux inconnues ;-)

La droite centrale représente le « biais » (terminologie employée dans l'article de Bland et Altman[10]) : la moyenne des écarts entre les deux séries. Sur la figure 6, on peut lire que l'ordonnée de la droite centrale (droite « biais ») vaut 36,7 µmol/L. Ainsi, comme indiqué par les auteurs dans la discussion leur article (cf. ci-dessous), la méthode de mesure « Hand-held meter » de la concentration en BHBA sur-estime, en moyenne, la méthode de mesure « Laboratory » de 36,7 µmol/L :


taken as gold standard (Fig. 1A and B). In fact, the hand-held meter overestimated BHBA concentrations (bias = +36.7 or expressed as a percentage, 4.4%); however, only one outlier was recorded, where

Les deux droites « extérieures » représentent les limites d'agrément (« limits of agreement ») inférieure et supérieure. Sur la figure 6, on peut lire que les ordonnées de ces deux droites valent respectivement -224,7 µmol/L et +298,1 µmol/L. L'interprétation de ces valeurs est la suivante : si l'ensemble des écarts entre les deux séries de mesures suit une loi parfaitement normale[11], 95% des écarts entre les mesures réalisées par la 2ème méthode et celles réalisées par la 1ère méthode sont à l'intérieur de l'intervalle {-224,7 µmol/L ; +298,1 µmol/L}.

3. *Calculs pour dresser un graphique de Bland et Altman*

Tout d'abord, première remarque avant de vous présenter les calculs pour dresser un graphique de Bland et Altman : les calculs sont basés sur l'hypothèse (très souvent vérifiée) que les écarts entre les deux séries de mesures suivent une loi normale. (Notons qu'il n'est pas nécessaire que les mesures elles-mêmes suivent une loi normale (Bland and Altman, 1999).)

Comme nous l'avons vu sur la figure 6, en plus des deux séries de mesures, nous avons besoin de créer deux variables supplémentaires : une variable valant la moyenne des deux valeurs pour chaque individu (une pour la série n°1 et une pour la série n°2 ; variable « MOYENNES »), et une variable valant la différence des deux valeurs pour chaque individu (variable « ECARTS »). Les points du graphique de Bland et Altman seront les points d'abscisse MOYENNES et d'ordonnée ECARTS, chaque point représentant un individu mesuré deux fois.

L'ordonnée de la droite « biais » (⇔ la valeur du « biais ») est la moyenne de la variable ECARTS (soit M_{ECARTS} cette valeur) calculée sur l'ensemble des individus. Pour calculer les ordonnées des limites d'agrément, il est nécessaire de calculer d'abord la « Standard Deviation » de la variable ECARTS sur l'ensemble des individus (soit SD_{ECARTS} cette valeur). L'ordonnée de la droite représentant la limite supérieure d'agrément (⇔ la valeur de la limite supérieure d'agrément) vaut $M_{ECARTS} + 1{,}96 \times SD_{ECARTS}$ et celle de la droite représentant la limite inférieure d'agrément vaut $M_{ECARTS} - 1{,}96 \times SD_{ECARTS}$. Ainsi, 95% des écarts sont compris dans l'intervalle $\{M_{ECARTS} \pm 1{,}96 \times SD_{ECARTS}\}$. Si vous commencez à regarder dans vos données s'il y a effectivement 95% des écarts compris dans cet intervalle calculé, vous avez de bonnes chances de ne pas en trouver pile 95%, mais un peu plus ou un peu moins. Ne vous affolez pas, c'est parce que vos écarts entre les deux séries de mesures ne suivent pas une loi parfaitement normale. Dans le doute, n'hésitez pas à dresser un histogramme de l'ensemble des écarts entre les deux séries de mesures.

[10] Je ne défends pas cette terminologie, car un « biais » est un écart systématique entre une estimation et la valeur correspondante dans la population que l'on vise (par exemple, l'écart systématique entre l'estimation d'un taux de prévalence d'une maladie dans un échantillon et la valeur de ce taux de prévalence dans la population cible).

[11] C'est une hypothèse que l'on fera toujours pour interpréter un graphique de Bland et Altman, et que l'on pourrait néanmoins systématiquement vérifier à l'aide d'un histogramme.

4. *Inférence en utilisant la méthode de Bland et Altman*

Comme pour toute estimation, le « biais » estimé et les limites d'agrément estimées nécessitent leur intervalle de confiance à 95% avant de faire de l'inférence (c'est-à-dire, avant de généraliser les résultats de l'étude à la population cible) (Sim and Reid, 1999). L'article de Bland et Altman publié en 1999 fournit les formules pour calculer l'intervalle de confiance des limites d'agrément (Bland and Altman, 1999). Ces intervalles de confiance devraient être fournis au moment de fournir les limites d'agrément afin d'évaluer le degré d'incertitude lors de l'estimation des limites d'agrément, et d'inférer correctement ces limites d'agrément. Cependant, les formules fournies par Bland et Altman dans leur article de 1999 ont été remises en question, et je suggère la lecture de l'article de Carkett et Goh si ce point vous intéresse (Carkeet and Goh, 2018). Les intervalles de confiance des limites d'agrément calculés et fournis dans le ficher Excel® ont été calculés à partir des formules fournies par Bland et Altman dans leur article de 1999.

5. *Présentation des données pour les exemples*

L'un des objectifs du travail de la thèse vétérinaire du Dr C Boyer était d'évaluer la reproductibilité inter-opérateurs d'un protocole de mesure de dimensions de la rate du Gris du Gabon, à partir de clichés radiographiques. Deux opérateurs ont effectué les mesures : le Dr M Huynh (MH) et le Dr C Boyer (CB). Les dimensions de la rate sur lesquelles nous allons nous focaliser sont (cf. figure 7) :

- La hauteur dorso-ventrale maximale du bréchet perpendiculaire à la carène ventrale du bréchet, en mm (HDV) ;

- La hauteur maximale de la cavité coelomique en passant par le centre de la rate, en mm (HCC).

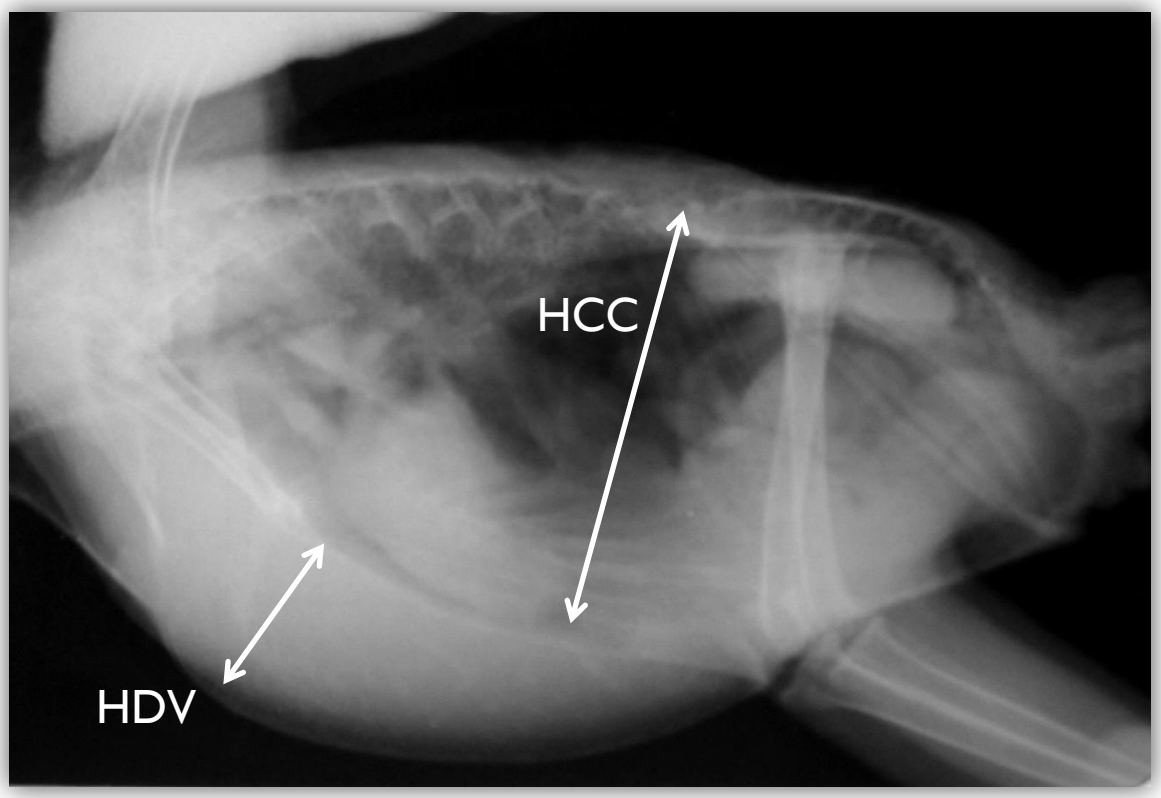
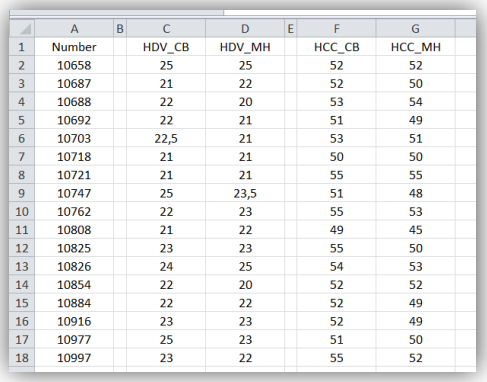


Figure 7. Radiographie d'un Gris du Gabon mettant en évidence les deux mesures qui vont être traitées dans ce guide pratique pour illustrer la méthode de Bland et Altman : HDV (hauteur dorso-ventrale maximale du bréchet) et HCC (hauteur maximale de la cavité coelomique).

Les mesures de HDV et HCC ont été réalisées par MH et par CB sur 41 Gris du Gabon (chaque cliché radiographique d'un Gris du Gabon a été évalué par MH et par CB ; le cliché radiographique est donc l' « individu »). La figure ci-dessous présente un extrait des mesures réalisées par les deux opérateurs pour chacun des 41 Gris du Gabon.



	A	C	D	F	G
1	Number	HDV_CB	HDV_MH	HCC_CB	HCC_MH
2	10658	25	25	52	52
3	10687	21	22	52	50
4	10688	22	20	53	54
5	10692	22	21	51	49
6	10703	22,5	21	53	51
7	10718	21	21	50	50
8	10721	21	21	55	55
9	10747	25	23,5	51	48
10	10762	22	23	55	53
11	10808	21	22	49	45
12	10825	23	23	55	50
13	10826	24	25	54	53
14	10854	22	20	52	52
15	10884	22	22	52	49
16	10916	23	23	52	49
17	10977	25	23	51	50
18	10997	23	22	55	52

Figure 8. Extrait des données recueillies par CB pour sa thèse vétérinaire. « CB » pour l'opérateur C Boyer, et « MH » pour l'opérateur M Huynh. « Number » fait référence au numéro d'identifiant unique de chaque Gris du Gabon évalué. La figure présente les données des 17 premiers Gris du Gabon évalués.

6. *Etape indispensable à réaliser avant de dresser un graphique de Bland et Altman*

Avant d'utiliser la méthode graphique de Bland et Altman sur vos données, il est absolument indispensable d'avoir une idée *a priori* des valeurs X et Y dans la phrase suivante, en fonction des 4 contextes présentés au début de ce guide (cf. tableau 1). Les valeurs de X et Y sont les 2 critères de Bland et Altman qu'il faut fixer *a priori* (c'est-à-dire, avant d'avoir recueilli vos données, ou à minima, avant de *regarder* vos données) et qui doivent être vérifiés dans l'étude pour pouvoir considérer que la concordance est satisfaisante (Chhapola et al., 2015; Mantha et al., 2000).

Contexte n°1 (répétabilité d'une méthode de mesure) : « je considère que la méthode de mesure est répétable si (1) en moyenne, la deuxième série de mesures ne sur-estime ou ne sous-estime pas les valeurs par rapport à la première série de mesures de plus de X, et si (2) la quasi-totalité des écarts (en valeur absolue) entre les deux séries de mesures est inférieure à Y ». Dans ce contexte n°1 (et seulement dans ce contexte), il est possible que vous ne jugiez pas important le critère X. En revanche, le critère Y est *le* critère qui doit être vérifié pour garantir la répétabilité de votre méthode de mesure.

Contexte n°2 (reproductibilité inter-opérateurs d'une méthode de mesure) : « je considère que les deux opérateurs donnent des valeurs concordantes si (1) en moyenne, l'un ne sur-estime ou ne sous-estime pas les valeurs par rapport à l'autre opérateur de plus de X, et si (2) la quasi-totalité des écarts (en valeur absolue) entre les deux opérateurs est inférieure à Y ».

Contexte n°3 (reproductibilité spacio-temporelle d'une méthode de mesure) : « je considère que la méthode de mesure est reproductible si (1) en moyenne, la deuxième série de mesures ne sur-estime ou ne sous-estime pas les valeurs par rapport à la première série de mesures de plus de X, et si (2) la quasi-totalité des écarts (en valeur absolue) entre les deux séries de mesures est inférieure à Y ».

Contexte n°4 (concordance entre deux méthodes de mesure) : « je considère que les deux méthodes de mesure sont concordantes si (1) en moyenne, la deuxième méthode de mesure ne sur-estime ou ne sous-estime pas les valeurs par rapport à la première méthode de mesure de plus de X, et si (2) la quasi-totalité des écarts (en valeur absolue) entre les deux séries de mesures provenant des deux méthodes de mesure est inférieure à Y ».

La notion de « quasi-totalité des écarts » pour le 2ème critère correspond à « 95% des écarts » dans la méthode de Bland et Altman. Mais j'ai volontairement écrit « quasi-totalité des écarts » plutôt que « 95% des écarts » car c'est dans ces termes que vous devez vous poser la question pour fixer la valeur de Y. Cette « quasi-totalité » doit être comprise de la façon suivante : « tous les écarts à part les quelques (5%) écarts les plus extrêmes ».

Une fois les valeurs X et Y des premier et second critères fixées *a priori*, il faudra vérifier que vos données respectent ces deux critères. Le premier critère fait référence au « biais » de Bland et Altman. Ainsi, il faudra vérifier que la moyenne des écarts observée dans l'échantillon (« biais » observé) est inférieure ou égale à la valeur X fixée *a priori*, que l'on peut donc appeler « biais maxi acceptable ». Le second critère fait référence à la zone d'agrément, la zone dans laquelle se trouvent 95% des écarts, définie par les limites inférieures et supérieures d'agrément. Ainsi, il faudra vérifier que la zone d'agrément estimée dans l'échantillon est incluse dans la zone {-Y ; +Y}, que l'on peut appeler « zone d'agrément maxi acceptable » (Chhapola et al., 2015).

Appliquons maintenant tout cela aux données sur les Gris du Gabon. Dans la thèse de CB, les valeurs de X et de Y avaient été fixées *a priori* aux valeurs présentées dans le tableau 11 pour les dimensions HDV et HCC.

Tableau 11. Valeurs des critères de Bland et Altman fixées *a priori* pour la hauteur dorso-ventrale maximale du bréchet (HDV) et la hauteur maximale de la cavité coelomique (HCC) des 41 Gris du Gabon évalués.

Critères de Bland et Altman	HDV	HCC
1er critère, biais maxi acceptable (X)	1 mm	2,5 mm
2ème critère, zone d'agrément maxi acceptable (±Y)	±2 mm	±5 mm

L'interprétation des données du tableau 11 pour la colonne HDV est la suivante : avant de commencer les mesures à partir des radiographies, CB avait considéré que la méthode de mesure de la dimension HDV était reproductible d'un opérateur à l'autre si (1) en moyenne, l'un ne sur-estimait ou ne sous-estimait pas les valeurs de HDV par rapport à l'autre opérateur de plus de 1 mm (« biais maxi acceptable »), et si (2) la quasi-totalité (95%) des écarts entre les deux opérateurs étaient compris dans la zone {-2 mm ; +2 mm}.

Le tableau 12 ci-dessous présente un extrait des 8 premières données des deux séries de mesures de la dimension HDV (la série de 8 mesures par CB et la série de 8 mesures par MH), ainsi que la colonne « MOYENNES » (qui est la moyenne, pour chaque Gris du Gabon évalué, de la valeur fournie par CB et de celle fournie par MH) et la colonne « ECARTS » (qui est la différence entre la valeur fournie par MH et celle fournie par CB). (Comme je l'ai dit précédemment, ce sont ces colonnes « MOYENNES » et « ECARTS » qui vont être utilisées pour dresser le graphique de Bland et Altman.)

Tableau 12. Extrait des valeurs de HDV mesurées (en mm) par CB et par MH.

N° du Gris du Gabon	HDV_CB	HDV_MH	MOYENNES	ECARTS
10658	25	25	25	0
10687	21	22	21,5	1
10688	22	20	21	-2
10692	22	21	21,5	-1
10703	22,5	21	21,75	-1,5
10718	21	21	21	0
10721	21	21	21	0
10747	25	23,5	24,25	-1,5

La figure 9 correspond au graphique de Bland et Altman, fourni par le fichier Excel®, évaluant la concordance inter-opérateurs dans les mesures de HDV.

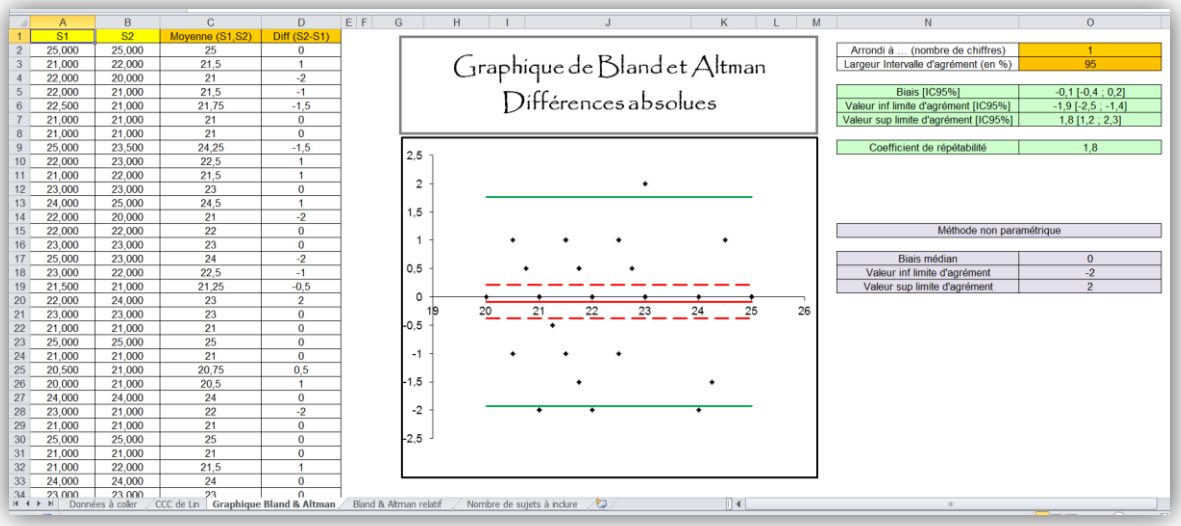


Figure 9. Graphique de Bland et Altman représentant la concordance des mesures de HDV entre celles effectuées par CB et celles effectuées par MH.

Dans la figure 9, chaque point représente 1 Gris du Gabon évalué (il y a donc 41 points sur la figure[12]). Chaque point a pour abscisse la valeur de la moyenne des deux mesures réalisées sur le Gris du Gabon par MH et CB (colonne « MOYENNES » dans le tableau 12 ; colonne « C » dans la figure 9), et pour ordonnée la différence entre la mesure de MH et celle de CB (colonne « ECARTS » dans le tableau 12 ; colonne « D » dans la figure 9).

La droite rouge (quasiment sur l'axe des abscisses) représente le « biais » observé, c'est-à-dire la moyenne des écarts de mesures entre MH et CB. Ici, la valeur de ce « biais » est -0,1 (cf. en haut à droite de la figure 9). Cela signifie qu'en moyenne, MH (la série n°2) sous-estime la dimension HDV de 0,1 mm par rapport à CB (la série n°1). (Si la valeur de ce « biais » avait été de +0,1 mm, cela aurait signifié qu'en moyenne, MH aurait *sur*-estimé la dimension HDV de 0,1 mm par rapport à CB.) L'intervalle de confiance à 95% de ce « biais » vaut [-0,4 mm ; +0,2 mm] (cf. en haut à droite de la figure 9). Cet intervalle de confiance à 95% comprend la valeur « 0 », donc il n'y a pas de sur-estimation ou de sous-estimation significative de la façon de mesurer la longueur HDV entre les deux opérateurs. Le 1er critère de Bland et Altman pour HDV (cf. tableau 11) était de 1 mm. Ce critère est donc rempli, puisque MH a sous-estimé en moyenne la mesure HDV de seulement 0,1 mm par rapport à CB (< 1 mm fixé *a priori*).

Les deux droites vertes représentent les limites d'agrément à 95%. Elles ont pour ordonnées -1,9 mm et +1,8 mm, respectivement pour les limites d'agrément inférieure et supérieure (cf. nombres fournis en haut à droite de la figure 9). Cette zone d'agrément estimée, définie par les valeurs de ces limites d'agrément à 95%, s'interprète de la façon suivante : la méthode de Bland et Altman estime que 95% des 41 écarts de mesures entre CB et MH sont compris entre -1,9 mm et +1,8 mm[13]. Le 2ème critère de Bland et Altman pour HDV (cf. tableau 11) était de ±2 mm. Ce critère est donc par conséquent lui aussi vérifié (c'était juste !), car ce critère imposait que 95% des écarts soient inclus à l'intérieur de la zone d'agrément. Pour information, le coefficient de concordance de Lin vaut 0,79 (niveau

[12] Si vous les comptez, vous en verrez bien moins que 41. C'est normal, il y a beaucoup d' « ex-aequo » (points ayant les mêmes abscisse et ordonnée).

[13] sous l'hypothèses que l'ensemble des écarts de mesures entre CB et MH suive une loi normale (colonne « D » de la figure 9).

« satisfactory » selon l'interprétation dans le tableau 10, et « substantial » d'après Landis et Koch dans le tableau 4).

La figure 9 montre aussi que la feuille Excel® fournit le « biais » observé ainsi que les limites d'agrément calculés de façon non paramétrique (c'est-à-dire, en calculant la médiane, le 2,5^{ème} percentile et le 97,5^{ème} percentile des écarts). Ces informations ne sont fournies qu'à titre illustratif, car la méthode de Bland et Altman utilise seulement la méthode paramétrique faisant l'hypothèse de la normalité des écarts entre les deux séries de mesures. Cependant, l'avantage de ces estimations non paramétriques du « biais » observé et des limites d'agrément est que ces estimations ne reposent pas sur l'hypothèse de la normalité des écarts. Vous trouverez plus d'informations sur la définition non paramétrique du « biais » observé et de la zone d'agrément dans l'article de Twomey de 2006 (Twomey, 2006).

Passons maintenant à la dimension HCC. Le tableau 13 ci-dessous présente un extrait des 8 premières séries de mesures de HCC, dont les colonnes ont la même interprétation que celles du tableau 12.

Tableau 13. Extrait des valeurs de HCC mesurées (en mm) par CB et par MH.

N° du Gris du Gabon	HCC_CB	HCC_MH	MOYENNES	ECARTS
10658	52	52	52	0
10687	52	50	51	-2
10688	53	54	53,5	1
10692	51	49	50	-2
10703	53	51	52	-2
10718	50	50	50	0
10721	55	55	55	0
10747	51	48	49,5	-3

La figure 10 correspond au graphique de Bland et Altman évaluant la concordance inter-opérateurs dans les mesures de HCC, fourni par le fichier Excel®.

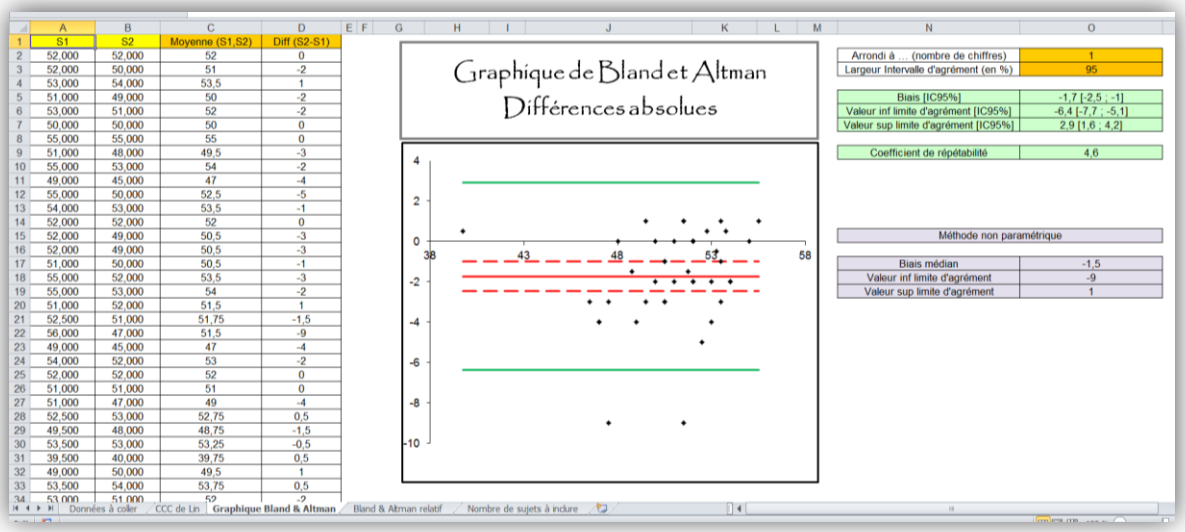


Figure 10. Graphique de Bland et Altman représentant la concordance des mesures de HCC entre celles effectuées par CB et celles effectuées par MH.

Sur cette figure 10, on peut lire la valeur du « biais » : -1,7 mm. Cela signifie que MH sous-estime en moyenne de 1,7 mm la dimension HCC par rapport à CB. Ce « biais » (écart moyen) est significatif[14] puisque son intervalle de confiance à 95% (-2,5 ; -1,0) ne comprend pas la valeur « 0 ». Regardons cependant si le 1er critère de concordance est vérifié (cf. tableau 11) : X=2,5 mm. Par conséquent, le 1er critère est (quand même) vérifié car le « biais » observé de 1,7 mm reste inférieur à X=2,5 mm.

Les limites d'agrément à 95% valent -6,4 mm et +2,9 mm. Cela signifie que la quasi-totalité (95%) des écarts de mesures entre CB et MH sont inclus dans la zone {-6,4 mm ; 2,9 mm}. Or, le 2ème critère fixé *a priori* était Y=±5 mm (cf. tableau 11). Par conséquent, le 2ème critère, lui, n'est pas vérifié car la zone d'agrément observée n'est pas incluse à l'intérieur de la zone d'agrément maxi acceptable. Pour information, le coefficient de concordance de Lin vaut 0,62 (niveau « mediocre » selon l'interprétation dans le tableau 10, mais encore « substantial » d'après Landis et Koch dans le tableau 4).

La conclusion de cette étude sur la reproductibilité inter-opérateurs de la mesure de deux dimensions de la rate du Gris du Gabon HDV et HCC serait que le protocole de lecture de la dimension HDV serait acceptable car reproductible, mais le protocole de lecture de la dimension HCC ne remplit pas un des deux critères de reproductibilité, et n'est donc par conséquent pas acceptable.

7. *Le coefficient de répétabilité de la méthode graphique de Bland et Altman*

Bland et Altman définissent aussi un « coefficient de répétabilité » (Bland and Altman, 1999). Il a pour valeur $1{,}96 \times SD_{ECARTS}$. Sachant que la zone d'agrément est définie par $\{M_{ECARTS} \pm 1{,}96 \times SD_{ECARTS}\}$, le coefficient de répétabilité vaut la moitié de la largeur de la zone d'agrément[15]. Sur la figure 10 ci-dessus, vous pouvez voir que la zone d'agrément est {-6,4 ; 2,9}. La largeur de cette zone est donc : 2,9 – (-6,4) = 9,3. La moitié de cette largeur vaut : 9,3/2 = 4,65, qui est justement la valeur du coefficient de répétabilité que vous pouvez voir en haut à droite de la figure 10 (aux arrondis près). L'interprétation de la valeur du coefficient de répétabilité est la suivante : 95% des écarts entre les deux séries de mesures, en valeur absolue, sont inférieurs ou égaux à la valeur du coefficient de répétabilité, *en retirant l'écart systématique (le « biais » observé) entre les deux séries de mesures*. C'est cette dernière partie de la phrase en italique qui rend l'interprétation du coefficient de répétabilité difficile : deux séries de mesures sont concordantes globalement, ou ne le sont pas, et il serait difficile de dire « elles le sont, si l'on retire (ou si l'on fait abstraction) de l'écart systématique entre les deux séries de mesures ». Cependant, il existe une (au moins, à ma connaissance) situation où le coefficient de répétabilité pourrait être malgré tout pertinent. Supposons la situation de concordance entre deux méthodes de mesure, avec une méthode de référence difficilement réalisable sur le terrain et une nouvelle méthode beaucoup plus facilement réalisable sur le terrain. Un « biais » non nul et cliniquement important conduirait à penser qu'il faudrait corriger les valeurs fournies par la nouvelle méthode en ajoutant (ou en retranchant) systématiquement la valeur du « biais » estimée (sous réserve que le « biais » estimé soit estimé avec précision[16]). Une fois ces valeurs issues de la nouvelle méthode corrigées, la droite rouge (le « biais ») serait donc confondue avec la droite d'ordonnée = 0, et les limites d'agréments à 95% seraient {- coefficient de répétabilité ; + coefficient de répétabilité}. Par conséquent, la valeur du coefficient de répétabilité peut s'interpréter de la façon suivante : il s'agit de la zone d'agrément si l'on avait appliqué la mesure correctrice afin de rendre les valeurs de la nouvelle méthode égale *en moyenne* aux valeurs de la

[14] C'est-à-dire, significativement différent de « 0 ».

[15] La question de la pertinence du mot « coefficient » choisi par Bland et Altman se pose. En effet, un « coefficient » est censé être sans dimension, alors que la valeur de 1,96xSD a la dimension des mesures : des mm, g/L, kg, *etc*. (Patton et al., Eye, 2006).

[16] Avec un intervalle de confiance à 95% resserré autour de la valeur du biais, intervalle fourni par le fichier Excel® en haut à droite (par exemple : [-2,48 ; -0,99] dans la figure 10).

méthode de référence (en retranchant le « biais » aux valeurs de la nouvelle méthode ; après « étalonnage », en quelque sorte).

Attention donc à ne pas sur-interpréter un coefficient de répétabilité faible. La figure 11 ci-dessous illustre un exemple où le coefficient de répétabilité est relativement faible, mais les deux séries ne sont pas concordantes (car l'une sur-estime de beaucoup l'autre).

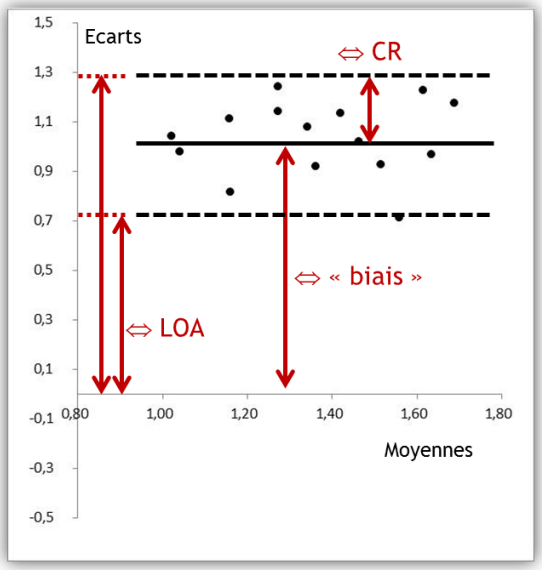


Figure 11. Illustration de deux séries non concordantes car le « biais » est important, malgré un coefficient de répétabilité (CR) faible. Les limites d'agrément (LOA) montrent aussi cette mauvaise concordance avec une limite supérieure (1,3) élevée.

8. *La méthode de Bland et Altman avec les différences relatives*

Comme nous venons de le voir, le graphique original de Bland et Altman représente en ordonnées des différences absolues entre les valeurs fournies par la première série de mesures et celles fournies par la seconde. Mais une autre méthode a été proposée, en plaçant en ordonnées des différences *relatives* exprimées en %, entre les deux séries de mesures (Pollock et al., 1992). Je recommande fortement la lecture de l'article de Twomey qui présente les avantages et les inconvénients des différentes méthodes graphiques de Bland et Altman (différences absolues ou relatives en ordonnées) (Twomey, 2006).

La méthode originale (différences absolues) ne devrait pas être utilisée lorsque, visuellement, les écarts en valeur absolue augmentent avec la valeur mesurée (ensemble des points formant un cône fermé à gauche et ouvert à droite sur le graphique de Bland et Altman). La méthode des différences relatives devrait être utilisée dans cette situation-là. (La transformation logarithmique a par ailleurs été proposée dans le cas où la méthode des différences absolues ne peut pas être utilisée (Hollis, 1996).) Dans la mesure où cette méthode des différences relatives fournies des écarts entre les deux séries de mesures en %, il peut être plus facile d'utiliser cette méthode plutôt que celle des différences absolues. Cependant, la méthode des différences relatives ne devraient pas être utilisée si le graphique utilisant cette méthode représente un ensemble de points formant un cône ouvert à gauche et fermé à droite.

Sur les figures 9 et 10, les nuages de points ne forment pas de cône fermé à gauche et ouvert à droite, donc la méthode des différences absolues pouvait tout à fait être utilisée.

La méthode des différences relatives place chaque point avec pour abscisse la moyenne des deux valeurs, et en ordonnées, la différence des deux valeurs *rapportées* à la moyenne de ces deux valeurs : (série 2 – série 1) / ((série 1 + série 2)/2).

Le tableau 14 ci-dessous présente les données du tableau 12, avec la colonne supplémentaire « ECARTS RELATIFS » valant, pour chaque Gris du Gabon, la différence entre la mesure de MH et celle de CB divisée par la moyenne des deux mesures (différence exprimée en %). Ce sont les colonnes « MOYENNES » et « ECARTS RELATIFS » qui vont être utilisées pour dresser le graphique de Bland et Altman avec différences relatives.

Tableau 14. Extrait des valeurs de HDV mesurées (en mm) par CB et par MH.

N° du Gris du Gabon	HDV_CB	HDV_MH	MOYENNES	ECARTS	ECARTS RELATIFS
10658	25	25	25	0	0,0%
10687	21	22	21,5	1	4,7%
10688	22	20	21	-2	-9,5%
10692	22	21	21,5	-1	-4,7%
10703	22,5	21	21,75	-1,5	-6,9%
10718	21	21	21	0	0,0%
10721	21	21	21	0	0,0%
10747	25	23,5	24,25	-1,5	-6,2%

A titre d'illustration, la valeur de « -6,9% » du Gris du Gabon n°10703 dans le tableau 14 a été obtenue de la façon suivante : (21-22,5) / ((21+22,5)/2) = (21-22,5) / 21,75 = -0,069 = -6,9%.

La méthode de Bland et Altman avec différences relatives demande de fixer *a priori* les valeurs des critères X et Y non plus en valeur absolue, mais en valeur relative (exprimée en %). Ces valeurs peuvent être difficiles à anticiper, puisque la différence entre les deux séries est divisée par quelque chose de difficile à appréhender : la moyenne des valeurs issues des deux séries. La tableau 15 ci-dessous peut aider à fixer les critères de Bland et Altman X et Y en valeurs relatives.

Tableau 15. Valeur du coefficient multiplicatif en fonction des valeurs de la différence relative (en valeur absolue) entre deux séries de mesures.

|Différence relative|	Coefficient multiplicatif
0%	1,00
1%	1,01
2%	1,02
3%	1,03
4%	1,04
5%	1,05
6%	1,06
7%	1,07
8%	1,08
9%	1,09
10%	1,11
20%	1,22
30%	1,35
50%	1,67
80%	2,33
100%	3,00

La tableau 15 se lit de la façon suivante, en prenant l'exemple d'une |différence relative| égale à 30% qui correspond à un coefficient multiplicatif de 1,35 : s'il existe une |différence relative| entre deux mesures (d'un même individu) de 30%, c'est que la plus élevée des deux est 1,35 fois plus grande que l'autre. On peut remarquer dans le tableau 15 que lorsque la |différence relative| est < 10%, l'interprétation est facile en 1ère approximation : une |différence relative| de V entre les deux séries de mesures indique que la plus élevée des deux est 1 + V fois plus grande que l'autre. Par

exemple, regardez dans le tableau 14 le Gris du Gabon n°10747. MH avait mesuré la dimension HDV avec 1,5 mm de moins que CB (23,5 *vers*us 25 mm). La différence absolue est de -1,5 mm, la |différence relative| est de 6,2%. A l'aide du tableau 15, on peut interpréter ce 6,2% (V vaut donc 0,062) : la mesure de CB était 1 + 0,062 ≈ 1,062 fois plus grande que la mesure de MH pour ce Gris du Gabon (en effet, 25/23,5 = 1,064, proche du 1,062 en approximation). La correspondance facile entre |différence relative| et coefficient multiplicatif devient moins évidente au fur et à mesure que la |différence relative| est éloignée de 10% par valeur supérieure…

Au moment de la thèse du Dr. C Boyer, le raisonnement n'avait porté que sur les critères « absolus » de Bland et Altman (cf. tableau 11). Supposons maintenant que l'on ait fixé *a priori* les critères de Bland et Altman X et Y de façon relative, tel qu'indiqué dans le tableau 16 ci-dessous.

Tableau 16. Valeurs des critères « relatifs » de Bland et Altman pour HDV et HCC des 41 Gris du Gabon évalués.

Critères de Bland et Altman	HDV	HCC
1er critère, biais relatif maxi acceptable (X)	5%	5%
2ème critère, zone d'agrément maxi acceptable (Y)	±10%	±10%

S'imposer les critères indiqués dans le tableau 16 revient à s'imposer les conditions suivantes :

- X relatif = 5% : en moyenne, l'un ne sur-estime ou ne sous-estime pas les valeurs de HDV par rapport à l'autre opérateur de plus de 5% de la moyenne des valeurs des deux opérateurs,
- Y relatif = 10% : 95% des écarts relatifs entre les deux opérateurs sont compris dans la zone {-10% ; +10%} de la moyenne des valeurs des deux opérateurs.

Le fichier Excel® dressant le graphique de Bland et Altman telles que celles représentées sur les figures 9 et 10 fournit aussi le graphique de Bland et Altman pour les différences relatives. La figure 12 présente le graphique de Bland et Altman en utilisant la méthode des différences relatives, pour les mesures de HDV effectuées par CB et MH. La figure 13 présente ce même graphique pour les mesures de HCC effectuées par CB et MH.

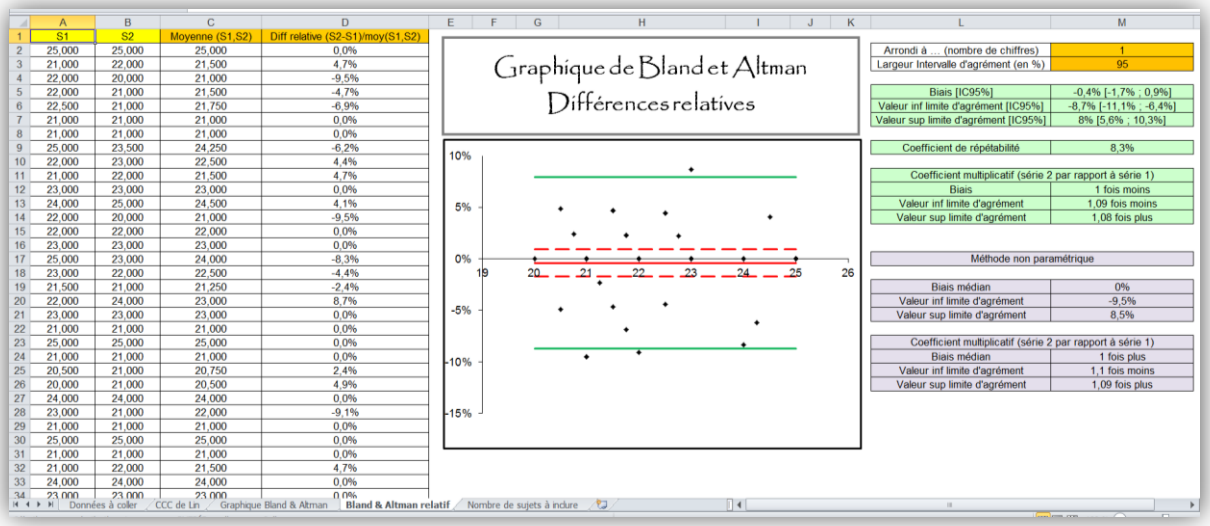


Figure 12. Graphique de Bland et Altman représentant la concordance des mesures de HDV entre celles effectuées par CB et celles effectuées par MH, en utilisant les différences relatives.

On peut tout d'abord remarquer que les points représentés sur la figure ne forment pas de cône ouvert à gauche et fermé à droite. Par conséquent, les valeurs du « biais » et des limites d'agrément, fournies par la méthode de Bland et Altman avec différences relatives, sont *a priori* interprétables. Sur la figure 12, la droite rouge d'ordonnées -0,4% (cf. en haut à droite de la figure) indique que le

deuxième opérateur (MH) sous-estime en moyenne les valeurs de HDV par rapport à l'autre opérateur (CB) de 0,4%. Cette valeur est inférieure à celle de 5% du tableau 16 ; par conséquent, le 1er critère de Bland et Altman serait vérifié. La zone d'agrément est {-8,7% ; +8,0%}. Cette zone d'agrément s'interprète de la façon suivante : la méthode de Bland et Altman estime que 95% des 41 écarts relatifs de mesures entre CB et MH sont compris entre -8,7% et +8,0%. Le 2ème critère de Bland et Altman pour HDV (cf. tableau 16) était de ±10%. Ce critère serait donc par conséquent lui aussi vérifié.

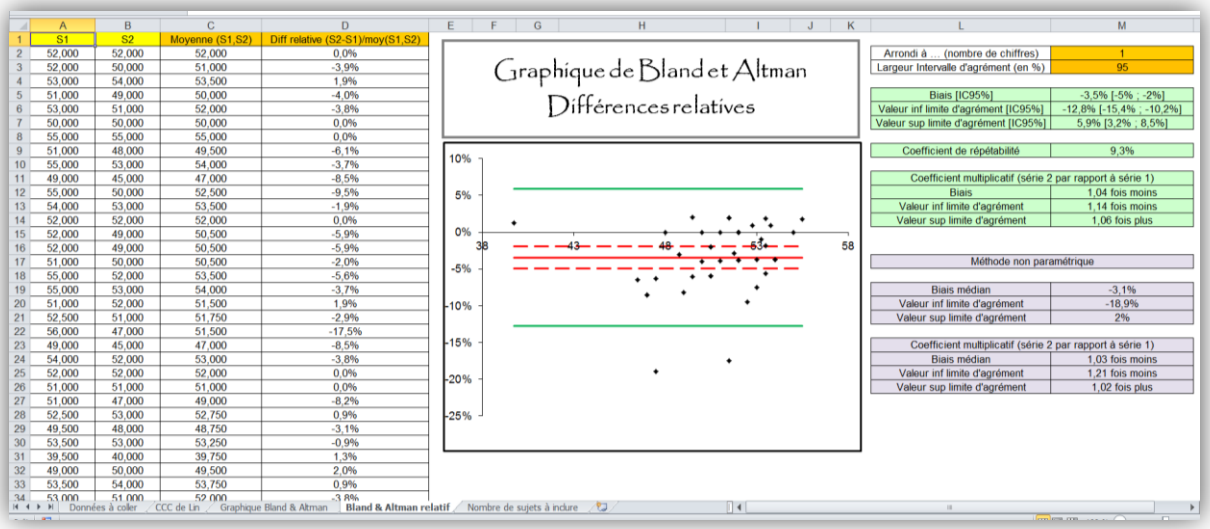


Figure 13. Graphique de Bland et Altman représentant la concordance des mesures de HCC entre celles effectuées par CB et celles effectuées par MH, en utilisant les différences relatives.

Sur la figure 13, la droite rouge d'ordonnées -3,5% (cf. en haut à droite de la figure) indique que le deuxième opérateur (MH) sous-estime en moyenne les valeurs de HCC par rapport à l'autre opérateur (CB) de 3,5%. Ce « biais » est significatif (car son intervalle de confiance à 95% ne contient pas « 0% »), mais cette valeur étant inférieure à celle de 5% du tableau 16, le 1er critère de Bland et Altman serait vérifié. La zone d'agrément est {-12,8% ; +5,9%}. Le 2ème critère de Bland et Altman pour HDV (cf. tableau 16) était de ±10%. Ce critère ne serait donc par conséquent pas vérifié.

On peut remarquer que, comme dans la méthode des différences absolues, le « biais » et les limites d'agrément sont aussi calculés de façon non paramétrique. Ces informations sont données à titre exclusivement informatif.

D. Confrontation entre coefficient de concordance de Lin et graphique de Bland et Altman

Comme on vient de le voir, le fait de considérer que deux séries de mesures sont concordantes de façon satisfaisante ou non à l'aide du graphique de Bland et Altman (vérification des deux critères X et Y) repose sur des considérations « cliniques » (choix *a priori* des valeurs X et Y). Dans le calcul du coefficient de concordance de Lin, la considération clinique est inexistante (le seuil que l'on se fixe pour considérer que la concordance est satisfaisante est fixé hors contexte – cf. tableaux 4 et 10). Or, même si l'on a parfois l'impression que certains reviewers de revues scientifiques préfèrent les considérations statistiques aux considérations cliniques, l'utilisation systématique du graphique de Bland et Altman pour évaluer la concordance de deux séries de mesures est vivement recommandée (Kottner et al., 2011) et cette méthode devrait donc être demandée par ces reviewers. En effet, un coefficient de concordance de Lin peut être élevé, avec des critères « cliniques » X et/ou Y cependant non vérifiés. A contrario, un coefficient de concordance de Lin peut révéler une concordance

considérée comme « moyenne » selon la classification de Landis et Koch (cf. tableau 4), avec des critères de Bland et Altman « cliniques » X et Y cependant tous deux vérifiés. Il ne faut pas voir cela comme une incohérence entre les deux méthodes d'évaluation de la concordance entre deux séries de mesures (statistique *versus* graphique/clinique) : les deux méthodes utilisent des critères différents pour juger de la répétabilité, reproductibilité, ou concordance de séries de mesures. A ce sujet, je ne résiste pas à l'envie de citer Bland et Altman dans leur article de 1990 (Bland and Altman, 1990) : « The magnitude of the difference [between two sets of measurements] which is acceptable is not a statistical decision, but a clinical one. We should ask whether the agreement is good enough for a particular purpose, not whether it conforms to some absolute, arbitrary criterion. Methods which may agree well enough for one purpose may not agree well enough for another. »

V. Coefficients de concordance, degré de signification, et inférence

Il est possible d'associer au coefficient de concordance (Kappa ou Lin) un degré de signification *p*. L'hypothèse nulle H_0 associée à ce degré de signification est l'absence de concordance réelle dans la population, entre deux séries de mesures. Ceci signifie que si $p < 0,05$, on fera l'inférence suivante : il y a de grandes chances pour que la méthode soit répétable, reproductible, ou concordante avec une autre. A première vue, il semble donc intéressant de tester statistiquement le coefficient de concordance pour savoir si son degré de signification est inférieur ou supérieur à 0,05, en espérant obtenir un $p < 0,05$. Cependant, quand on conçoit une étude pour quantifier la concordance entre deux séries de mesures, très souvent (sinon tout le temps), il existe réellement une concordance. Elle peut certes être faible, mais il serait difficilement concevable qu'elle soit *nulle*. Autrement dit, tester H_0 pour essayer de rejeter H_0 n'est pas pertinent car H_0 est très vraisemblablement fausse, et un non rejet de H_0 ($p > 0,05$) proviendrait bien davantage d'un manque de puissance statistique que du fait que H_0 soit vraiment vraie (« In a clinical setting, […] significance tests against zero are of little interest for such tests may show only that two methods agree more than by chance. It would be quite surprising if measurements obtained from two instruments designed to measure the same thing were not related. » (Muller and Buttner, 1994)).

Ainsi, ce que nous voulons montrer statistiquement n'est pas une concordance *non nulle*, mais montrer de façon statistique que la méthode de mesure est répétable/reproductible/concordante avec une autre méthode de façon *suffisamment* satisfaisante pour pouvoir la valider. Pour cela, il faut tout d'abord se fixer une valeur seuil minimale du coefficient de concordance à partir de laquelle on peut considérer la concordance comme « satisfaisante ». Ensuite, il faut montrer statistiquement que la valeur observée du coefficient de concordance est significativement supérieure à cette valeur seuil. Pour cela, il faut utiliser l'intervalle de confiance à 95% du coefficient de concordance estimé dans l'étude (Sim and Wright, 2005). Si la borne inférieure de l'intervalle de confiance est supérieure à cette valeur seuil minimale, alors vous aurez réussi à montrer statistiquement que la répétabilité/reproductibilité de la méthode ou la concordance entre les deux méthodes étudiées est significativement satisfaisante. Toute la question repose sur cette valeur seuil minimale. Si l'on se réfère au Tableau 4, une valeur supérieure à 0,60 pour un coefficient de concordance correspond à une concordance au moins « bonne » (« substantial »). Cette valeur pourra ainsi être considérée comme la valeur seuil minimale à partir de laquelle la concordance pourra être considérée comme « satisfaisante » (Donner and Eliasziw, 1992).

En reprenant l'exemple des diagnostics de réticulo-péritonite traumatique (RPT) réalisés par 2 vétérinaires sur 64 vaches examinées (cf. figure 1), la valeur du coefficient Kappa était de 0,75, et son degré de signification associé était < 0,01. L'intervalle de confiance à 95% du coefficient Kappa était [0,53 ; 0,88]. Cela signifie que si l'estimation du coefficient de concordance n'est pas biaisée, il y a 95% de chances pour que la vraie valeur du coefficient Kappa quantifiant la reproductibilité inter-opérateurs de la méthode de diagnostic de RPT soit comprise entre 0,53 et 0,88. Donc, on peut être confiant en pensant que la valeur réelle du coefficient Kappa est supérieure à 0,53. Sauf que 0,53 est

une valeur assez faible, n'atteignant pas la valeur seuil de 0,60. Ainsi, bien que l'estimation du coefficient Kappa laissait penser que la méthode de diagnostic de RPT pouvait être considérée comme « satisfaisante » (estimation du coefficient Kappa de 0,75 > 0,60 ; cf. tableau 4), cette étude n'a cependant pas réussi à le montrer de façon *significative*. Par conséquent, dire « le coefficient Kappa est significativement différent de « 0 »[17] donc la méthode est reproductible » est une sur-interprétation des résultats statistiques ! Certes, on peut rejeter l'hypothèse (nulle) (rejeter l'hypothèse que la méthode *n'*est *pas* reproductible), mais de là à dire que cette méthode devrait être utilisée car elle l'est, il y a un précipice que la valeur inférieure de l'intervalle de confiance à 95% du coefficient Kappa (0,53) empêche de franchir.

Evidemment, ce que je viens d'écrire ci-dessus pour le coefficient Kappa est directement transposable au coefficient de concordance de Lin.

VI. Protocole à mettre en place pour évaluer la répétabilité, reproductibilité, ou concordance de méthodes de mesure.

A. Démarche générale

Pour évaluer la répétabilité ou la reproductibilité d'une même méthode de mesure, ou encore la concordance entre deux méthodes de mesure, vous devez mesurer *deux* fois plusieurs individus. Le calcul du nombre d'individus à mesurer deux fois est présenté ci-dessous dans la partie B.

Il est tout à fait possible de quantifier la répétabilité ou la reproductibilité d'une même méthode de mesure à partir d'au moins trois mesures par individu, mais le traitement statistique est beaucoup plus compliqué que lorsque l'individu n'est mesuré que deux fois (notamment en utilisant de la modélisation linéaire à effets mixtes (Chetboul et al., 2004; Ferre et al., 2001)). Cela dit, deux mesures par individu sont considérées comme suffisantes pour évaluer la répétabilité / reproductibilité d'une méthode de mesure, ou la concordance entre deux méthodes de mesure (Walter et al., 1998). Ainsi, puisque le traitement statistique est beaucoup plus simple avec deux mesures par individu qu'avec trois mesures ou plus, je conseille vivement de ne mesurer un individu que deux fois pour évaluer la répétabilité ou reproductibilité d'une méthode de mesure, ou la concordance entre deux méthodes de mesure.

Le protocole doit être tel que vous devrez être capable de remplir le tableau 17. Si vous voulez évaluer la répétabilité d'une méthode de mesure, la série n°1 correspondra à la série mesurée à un instant t_1 et la série n°2 correspondra à la série mesurée un instant t_2 (« très faiblement » distant temporellement de t_1). Si vous voulez évaluer la reproductibilité inter-opérateurs d'une méthode de mesure, la série n°1 correspondra à l'opérateur n°1 et la série n°2 correspondra à l'opérateur n°2. Si vous voulez évaluer la reproductibilité spacio-temporelle d'une méthode de mesure, la série n°1 correspondra à la série mesurée à un instant t_1 et la série n°2 correspondra à la série mesurée un instant t_2 (« grandement » distant temporellement et/ou spatialement de t_1). Si enfin vous voulez évaluer la concordance entre deux méthodes de mesure différentes, la série n°1 correspondra à la méthode de mesure n°1, et la série n°2 correspondra à la méthode de mesure n°2.

[17] C'est vrai, car le degré de signification était < 0,05 (cf. figure 1).

Tableau 17. Tableau à remplir pour évaluer la répétabilité/reproductibilité/concordance de méthodes de mesures.

Individu	Série n°1	Série n°2
1	Mesure $M_{1\text{-}1}$	Mesure $M_{2\text{-}1}$
2	Mesure $M_{1\text{-}2}$	Mesure $M_{2\text{-}2}$
3	Mesure $M_{1\text{-}3}$	Mesure $M_{2\text{-}3}$
…	…	…

Dans le cas de l'évaluation de la reproductibilité inter-opérateurs, certains protocoles prévoient 3 opérateurs ou plus. Dans de tels cas, l'utilisation des méthodes simples présentées dans ce guide pratique conduit à évaluer la concordance entre les opérateurs n°1 et n°2, puis entre les opérateurs n°1 et n°3, puis entre les opérateurs n°2 et n°3, *etc*. Ceci rend difficile l'interprétation des résultats, et vous vous limiterez par conséquent en général à deux opérateurs pour une interprétation plus facile de l'étude de la reproductibilité inter-opérateurs. Cependant, il y a au moins une situation où l'inclusion de 3 opérateurs est pertinente : si les opérateurs n°1 et n°2 ont les mêmes compétences, et si l'opérateur n°3 a des compétences différentes de celles des deux premiers. Dans cette situation, quantifier la concordance entre les opérateurs n°1 et n°2 répond à la question suivante : « la méthode est-elle reproductible lorsqu'elle est utilisée par deux opérateurs ayant les mêmes compétences ? » (qui est la reproductibilité classique inter-opérateurs). Et quantifier la concordance entre les opérateurs n°1 (ou n°2) et n°3 répond à la question suivante : « la méthode est-elle reproductible lorsqu'elle est utilisée par deux opérateurs ayant des compétences différentes ? ».

Enfin, le choix de l'indicateur numérique de concordance dépend du type de mesure : (a) si le caractère mesuré est binaire ou qualitatif *nominal*, il faut utiliser le coefficient de concordance Kappa classique, (b) si le caractère mesuré est qualitatif *ordinal*, il faudrait utiliser le coefficient de concordance Kappa pondéré[18], (c) si le caractère mesuré est quantitatif, il faut utiliser le coefficient de concordance de Lin associé au graphique de Bland et Altman, en fixant *a priori* les valeurs des critères X et Y[19] de la méthode graphique de Bland et Altman.

B. Calcul du nombre d'individus à mesurer deux fois

1. *Principe de calcul*

Comme nous l'avons vu dans la partie V ci-dessus, un degré de signification $p < 0,05$ pour un coefficient de concordance n'apporte que très peu (voire pas) d'information : on peut uniquement en conclure que la vraie valeur du coefficient de concordance a de grandes chances d'être différente de 0. Nous avons vu que la démarche statistique repose sur le calcul de l'intervalle de confiance à 95% de l'estimation du coefficient de concordance, et la position de sa borne inférieure par rapport à une valeur seuil minimale de « bonne concordance » fixée *a priori*. En reprenant les valeurs du tableau 4, on peut considérer que la valeur de 0,60 est cette valeur minimale. Il faut donc concevoir votre étude afin de maximiser les chances que la borne inférieure de l'intervalle de confiance à 95% du coefficient de concordance (Kappa ou Lin) soit supérieure à cette valeur seuil. Autrement dit, il faut que votre étude soit suffisamment puissante statistiquement. Si elle l'est trop peu, elle aura donc peu de chances de vous permettre de conclure que la concordance des deux séries est « satisfaisante ». Les articles suivants vous permettront d'aller plus loin sur le sujet (Shieh, 2014; Sim and Wright, 2005; Zou, 2012).

[18] L'utilisation du coefficient de concordance Kappa classique est possible, mais le calcul ne prendra pas en compte l'aspect ordinal des données.

[19] C'est-à-dire, avant d'avoir vu les données ! Sinon, c'est trop tard, vous serez influencé(e) par vos données pour fixer X et Y !

2. *Cas où le caractère est binaire*

Le calcul du nombre minimal d'individus à mesurer deux fois repose sur des formules statistiques dont certaines peuvent être bien compliquées. L'article de Donner et Eliasziw (Donner and Eliasziw, 1992) fournit une formule relativement[20] simple :

$$N = \lambda(1, 1-\beta, \alpha)\left\{\frac{[\pi(1-\pi)(\kappa_1-\kappa_0)]^2}{\pi^2+\pi(1-\pi)\kappa_0} + \frac{2[\pi(1-\pi)(\kappa_1-\kappa_0)]^2}{\pi(1-\pi)(1-\kappa_0)} + \frac{[\pi(1-\pi)(\kappa_1-\kappa_0)]^2}{(1-\pi)^2+\pi(1-\pi)\kappa_0}\right\}^{-1}$$

Où :

π = proportion attendue de la présence du caractère dans l'échantillon (dans le tableau 3, la valeur observée de cette proportion parmi les 64 vaches évaluées par le vétérinaire n°1 était de 43/64=67% ; elle était de 44/64=69% parmi les mêmes 64 vaches évaluées par le vétérinaire n°2).

K0 = valeur seuil minimale du coefficient de concordance Kappa (choisir une valeur au moins égale à « 0,60 », correspondant au seuil de « bonne » concordance (cf. tableau 4)).

K1 = coefficient de concordance Kappa attendu dans l'échantillon avant de réaliser l'étude.

$\lambda(1, 1-\beta, \alpha)$ = valeur dépendant du risque d'erreur α de 1ère espèce (toujours fixée à 0,05), et du risque d'erreur de 2ème espèce β directement liée à la puissance statistique (puissance statistique = 1-β). Si l'on souhaite une puissance statistique de 80%, il faut fixer β à 0,20.

Le fichier Excel® calculant le coefficient de concordance Kappa permet de calculer le nombre d'individus nécessaires (à mesurer deux fois) en utilisant la formule ci-dessus, en fixant α à 0,05. La figure ci-dessous en donne un exemple :

Taux de succès attendu	0,40
Kappa mini H0 (K0)	0,60
Kappa attendu (K1)	0,90
Puissance	0,80
Taille échantillon	58

En utilisant les nombres fournis ci-dessus, si l'on s'attend à observer dans l'échantillon, globalement, environ 40% des individus avec le caractère présent, si l'on souhaite mettre en place une étude avec 80% de puissance pour montrer de façon significative au risque d'erreur α = 0,05 que la vraie valeur du coefficient Kappa est supérieure à 0,60, et si l'on pense que le coefficient de concordance Kappa attendu est de 0,90, il faudra mesurer deux fois 58 individus.

Admettons que vous n'ayez à votre disposition que 20 individus à mesurer deux fois, quelles sont les chances de montrer statistiquement que le coefficient de concordance Kappa réel est supérieur à 0,60, en faisant l'hypothèse qu'il y a 40% de présence du caractère, et que le coefficient Kappa attendu est de 0,90 ? Il faut alors tâtonner en modifiant le chiffre de la puissance pour arriver à une taille d'échantillon égale à 20 : 0,37 (cf. figure ci-dessous).

[20] Tout est effectivement relatif !

Taux de succès attendu	0,40
Kappa mini H0 (K0)	0,60
Kappa attendu (K1)	0,90
Puissance	0,37
Taille échantillon	20

Ainsi, avec 20 individus mesurés deux fois, et sous les hypothèses citées ci-dessus, l'étude n'aura que 37% de chances de montrer de façon significative, au risque d'erreur α=0,05, que la vraie valeur du coefficient Kappa de votre méthode de mesure est au moins égale à 0,60.

3. *Cas où le caractère est quantitatif*

Le principe de calcul rejoint celui du coefficient de concordance Kappa. Les formules permettant le calcul de ce nombre d'individus sont présentées entre autres dans l'article de Walter et coll. (Walter et al., 1998). Pour obtenir ce nombre d'individus, il faut fournir les informations suivantes :

CCC0 = valeur seuil minimale du coefficient de concordance de Lin (choisir une valeur au moins égale à « 0,60 », correspondant au seuil minimal de « bonne » concordance (cf. tableau 4)).

CCC1 = coefficient de concordance de Lin attendu dans l'échantillon avant de réaliser l'étude.

Puissance = souvent fixée à 80%.

Le fichier Excel® calculant le coefficient de concordance de Lin permet de calculer le nombre d'individus nécessaires (à mesurer deux fois) en utilisant la formule (12) de l'article de Walter et coll., en fixant α à 0,05. La figure ci-dessous en donne un exemple :

CCC Lin mini H0 (CCC0)	0,60
CCC Lin attendu (CCC1)	0,75
Puissance	0,80
Taille échantillon	81

Ainsi, si l'on souhaite mettre en place une étude avec 80% de chances de montrer de façon significative que le coefficient de concordance de Lin réel est supérieur à 0,60, et si l'on pense que le coefficient de concordance de Lin attendu est de 0,75, il faudra mesurer deux fois 81 individus.

C. Ne pas calculer de moyennes de mesures par individu pour valider une méthode de mesure

Il arrive parfois que les investigateurs d'une étude prévoient de mesurer deux fois ou plus un même individu. Si ces investigateurs mesurent seulement deux fois un individu avec une même méthode de mesure *et* s'ils utilisent les méthodes décrites dans ce guide pratique (ou bien entendu celles décrites dans les articles cités dans ce guide), alors c'est très bien. Mais ce que ces investigateurs ne doivent pas faire, c'est de calculer la moyenne, par individu, des k ($k \geq 2$) mesures réalisées avec la même méthode de mesure, pour ensuite utiliser cette moyenne pour chaque individu, pour, par exemple, évaluer la reproductibilité inter-opérateurs ou la concordance de méthodes de mesure.

En effet, une méthode est reproductible d'un opérateur à un autre si la *valeur* fournie par le premier opérateur pour un individu donné est jugée comme très voisine de la *valeur* fournie par le second opérateur pour le même individu donné. Une méthode ne peut pas être qualifiée de reproductible entre opérateurs si ce sont des *moyennes* de valeurs, pour un individu donné, qui sont jugées comme très voisines entres celles d'un premier et celle d'un second opérateur.

Par exemple, si un protocole de mesures prévoit que l'opérateur n°1 mesure trois fois N individus et que l'opérateur n°2 mesure aussi trois fois ces mêmes N individus, il ne sera pas question de calculer pour chacun des N individus la moyenne des trois mesures de l'opérateur n°1, et la moyenne des trois mesures de l'opérateur n°2. Parmi ces trois séries de mesures pour les opérateurs n°1 et n°2, il faudra n'en choisir … qu'une ! Je vous recommande de choisir la première série, car elle correspond *a priori* à celle la plus « naïve », et donc celle qui est la plus proche de celle qui sera réalisée par la suite à grande échelle. De façon générale, il faut choisir la série de mesures dont les conditions soient les plus proches de celles qui seront celles de l'utilisation de la méthode à grande échelle.

Par conséquent, avant de vous lancer dans un protocole, faites attention à ne pas mesurer un même individu plus que nécessaire !

VII. Références

Atkinson, G. and Nevill, A., 1997. Comment on the Use of Concordance Correlation to Assess the Agreement between Two Variables. Biometrics. 53, 775-7.

Bakker, J., Olree, M., Kaatee, R., de Lange, E.E., Moons, K.G., Beutler, J.J. and Beek, F.J., 1999. Renal volume measurements: accuracy and repeatability of US compared with that of MR imaging. Radiology. 211, 623-8.

Barnhart, H.X., Haber, M. and Song, J., 2002. Overall concordance correlation coefficient for evaluating agreement among multiple observers. Biometrics. 58, 1020-7.

Barnhart, H.X., Haber, M.J. and Lin, L.I., 2007. An overview on assessing agreement with continuous measurements. J Biopharm Stat. 17, 529-69.

Bartlett, J.W. and Frost, C., 2008. Reliability, repeatability and reproducibility: analysis of measurement errors in continuous variables. Ultrasound Obstet Gynecol. 31, 466-75.

Bergknut, N., Meij, B.P., Hagman, R., de Nies, K.S., Rutges, J.P., Smolders, L.A., Creemers, L.B., Lagerstedt, A.S., Hazewinkel, H.A. and Grinwis, G.C., 2013. Intervertebral disc disease in dogs - Part 1: A new histological grading scheme for classification of intervertebral disc degeneration in dogs. Vet J. 195, 156-63.

Bland, J.M. and Altman, D.G., 1986. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet. 1, 307-10.

Bland, J.M. and Altman, D.G., 1990. A note on the use of the intraclass correlation coefficient in the evaluation of agreement between two methods of measurement. Comput Biol Med. 20, 337-40.

Bland, J.M. and Altman, D.G., 1999. Measuring agreement in method comparison studies. Stat Methods Med Res. 8, 135-60.

Brenner, H. and Kliebsch, U., 1996. Dependence of weighted kappa coefficients on the number of categories. Epidemiology. 7, 199-202.

Carkeet, A. and Goh, Y.T., 2018. Confidence and coverage for Bland-Altman limits of agreement and their approximate confidence intervals. Stat Methods Med Res. 27, 1559-1574.

Carrasco, J.L. and Tover, L., 2003. Estimating the generalized concordance correlation coefficient through variance components. Biometrics. 59, 849-58.

Chen, C.-C. and Barnhart, H.X., 2008. Comparison of ICC and CCC for assessing agreement for data without and with replications. Comput Statist Data Anal. 53, 554-64.

Chetboul, V., Athanassiadis, N., Concordet, D., Nicolle, A., Tessier, D., Castagnet, M., Pouchelon, J.L. and Lefebvre, H.P., 2004. Observer-dependent variability of quantitative clinical endpoints: the example of canine echocardiography. J Vet Pharmacol Ther. 27, 49-56.

Chhapola, V., Kanwal, S.K. and Brar, R., 2015. Reporting standards for Bland-Altman agreement analysis in laboratory research: a cross-sectional survey of current practice. Ann Clin Biochem. 52, 382-6.

Cohen, J., 1960. A coefficient of agreement for nominal scales. Educ Psychol Meas. 20, 37-46.

Cohen, J., 1968. Weighted Kappa: nominal scale agreement with provision for scaled disagreement or partial credit. Psychol Bull. 70, 213-20.

de Vet, H.C., Terwee, C.B., Knol, D.L. and Bouter, L.M., 2006. When to use agreement versus reliability measures. J Clin Epidemiol. 59, 1033-9.

Donner, A. and Eliasziw, M., 1987. Sample size requirements for reliability studies. Stat Med. 6, 441-8.

Donner, A. and Eliasziw, M., 1992. A goodness-of-fit approach to inference procedures for the kappa statistic: confidence interval construction, significance-testing and sample size estimation. Stat Med. 11, 1511-9.

Durando, M.M., Corley, K.T., Boston, R.C. and Birks, E.K., 2008. Cardiac output determination by use of lithium dilution during exercise in horses. Am J Vet Res. 69, 1054-60.

Ferre, P.J., Concordet, D., Laroute, V., Chanoit, G.P., Ferre, J.P., Manesse, M. and Lefebvre, H.P., 2001. Comparison of ultrasonography and pharmacokinetic analysis of creatine kinase release for quantitative assessment of postinjection muscle damage in sheep. Am J Vet Res. 62, 1698-705.

Gibbons-Burgener, S.N., Kaneene, J.B., Lloyd, J.W., Leykam, J.F. and Erskine, R.J., 2001. Reliability of three bulk-tank antimicrobial residue detection assays used to test individual milk samples from cows with mild clinical mastitis. Am J Vet Res. 62, 1716-20.

Giori, L., Giordano, A., Giudice, C., Grieco, V. and Paltrinieri, S., 2011. Performances of different diagnostic tests for feline infectious peritonitis in challenging clinical cases. J Small Anim Pract. 52, 152-7.

Giraudeau, B. and Mary, J.Y., 2001. Planning a reproducibility study: how many subjects and how many replicates per subject for an expected width of the 95 per cent confidence interval of the intraclass correlation coefficient. Stat Med. 20, 3205-14.

Graham, P. and Jackson, R., 1993. The analysis of ordinal agreement data: beyond weighted kappa. J Clin Epidemiol. 46, 1055-62.

Hollis, S., 1996. Analysis of method comparison studies. Ann Clin Biochem. 33 (Pt 1), 1-4.

Kottner, J., Audige, L., Brorson, S., Donner, A., Gajewski, B.J., Hrobjartsson, A., Roberts, C., Shoukri, M. and Streiner, D.L., 2011. Guidelines for Reporting Reliability and Agreement Studies (GRRAS) were proposed. J Clin Epidemiol. 64, 96-106.

Kottner, J. and Streiner, D.L., 2011. The difference between reliability and agreement. J Clin Epidemiol. 64, 701-2; author reply 702.

Kraemer, C.H., Periyakoil, V.S. and Noda, A., 2002. Kappa coefficients in medical research. Stat Med. 21, 2109-29.

Krouwer, J.S. and Monti, K.L., 1995. A simple, graphical method to evaluate laboratory assays. Eur J Clin Chem Clin Biochem. 33, 525-7.

Lachin, J.M., 2004. The role of measurement reliability in clinical trials. Clin Trials. 1, 553-66.

Landis, J.R. and Koch, G.G., 1977. The Measurement of Observer Agreement for Categorical Data. Biometrics. 33, 159-74.

Lee, J., Koh, D. and Ong, C.N., 1989. Statistical evaluation of agreement between two methods for measuring a quantitative variable. Comput Biol Med. 19, 61-70.

Lin, L., Hedayat, A.S. and Wu, W., 2007. A unified approach for assessing agreement for continuous and categorical data. J Biopharm Stat. 17, 629-52.

Lin, L.I., 1989. A concordance correlation coefficient to evaluate reproducibility. Biometrics. 45, 255-68.

Lin, L.I., 2000. Total deviation index for measuring individual agreement with applications in laboratory performance and bioequivalence. Stat Med. 19, 255-70.

Ludbrook, J., 2002. Statistical techniques for comparing measurers and methods of measurement: a critical review. Clin Exp Pharmacol Physiol. 29, 527-36.

Luiz, R.R., Costa, A.J., Kale, P.L. and Werneck, G.L., 2003. Assessment of agreement of a quantitative variable: a new graphical approach. J Clin Epidemiol. 56, 963-7.

Maclure, M. and Willett, W.C., 1987. Misinterpretation and misuse of the kappa statistic. Am J Epidemiol. 126, 161-9.

Mantha, S., Roizen, M.F., Fleisher, L.A., Thisted, R. and Foss, J., 2000. Comparing methods of clinical measurement: reporting standards for bland and altman analysis. Anesth Analg. 90, 593-602.

Muller, R. and Buttner, P., 1994. A critical discussion of intraclass correlation coefficients. Stat Med. 13, 2465-76.

Nickerson, C.A.E., 1997. A Note On "A Concordance Correlation Coefficient to Evaluate Reproducibility". Biometrics. 53, 1503-7.

Norton, J.L., Nolen-Walston, R.D., Underwood, C., Slack, J., Boston, R. and Dallap, B.L., 2011. Comparison of water manometry to 2 commercial electronic pressure monitors for central venous pressure measurement in horses. J Vet Intern Med. 25, 303-6.

Partik, B.L., Stadler, A., Schamp, S., Koller, A., Voracek, M., Heinz, G. and Helbich, T.H., 2002. 3D versus 2D ultrasound: accuracy of volume measurement in human cadaver kidneys. Invest Radiol. 37, 489-95.

Patton, N., Aslam, T. and Murray, G., 2006. Statistical strategies to assess reliability in ophthalmology. Eye (Lond). 20, 749-54.

Perkins, J.D., Salz, R.O., Schumacher, J., Livesey, L., Piercy, R.J. and Barakzai, S.Z., 2009. Variability of resting endoscopic grading for assessment of recurrent laryngeal neuropathy in horses. Equine Vet J. 41, 342-6.

Pollock, M.A., Jefferson, S.G., Kane, J.W., Lomax, K., MacKinnon, G. and Winnard, C.B., 1992. Method comparison--a different approach. Ann Clin Biochem. 29 (Pt 5), 556-60.

Shieh, G., 2014. Sample size requirements for the design of reliability studies: precision consideration. Behav Res Methods. 46, 808-22.

Shrout, P.E. and Fleiss, J.L., 1979. Intraclass correlations: uses in assessing rater reliability. Psychol Bull. 86, 420-8.

Sim, J. and Reid, N., 1999. Statistical inference by confidence intervals: issues of interpretation and utilization. Phys Ther. 79, 186-95.

Sim, J. and Wright, C.C., 2005. The kappa statistic in reliability studies: use, interpretation, and sample size requirements. Phys Ther. 85, 257-68.

Tennent-Brown, B.S., Koenig, A., Williamson, L.H. and Boston, R.C., 2011. Comparison of three point-of-care blood glucose meters for use in adult and juvenile alpacas. J Am Vet Med Assoc. 239, 380-6.

Twomey, P.J., 2006. How to use difference plots in quantitative method comparison studies. Ann Clin Biochem. 43, 124-9.

Voyvoda, H. and Erdogan, H., 2010. Use of a hand-held meter for detecting subclinical ketosis in dairy cows. Res Vet Sci. 89, 344-51.

Walter, S.D., Eliasziw, M. and Donner, A., 1998. Sample size and optimal designs for reliability studies. Stat Med. 17, 101-10.

Watson, P.F. and Petrie, A., 2010. Method agreement analysis: a review of correct methodology. Theriogenology. 73, 1167-79.

Zou, G.Y., 2012. Sample size formulas for estimating intraclass correlation coefficients with precision and assurance. Stat Med. 31, 3972-81.