Aide à la lecture de l'article

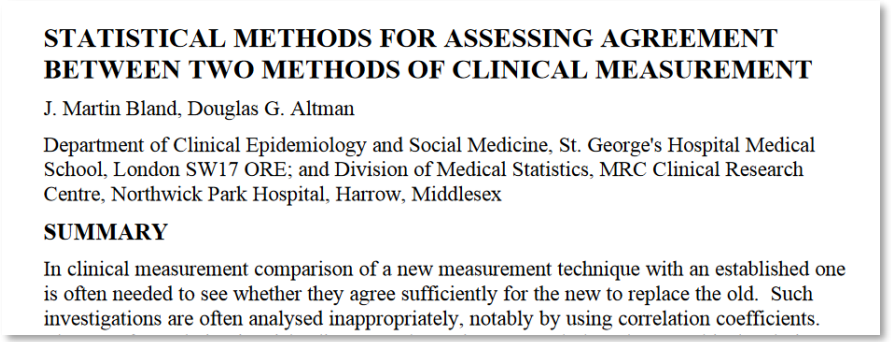
Bland, J. M., & Altman, D. G. (2010). **Statistical methods for assessing agreement between two methods of clinical measurement**. *International Journal of Nursing Studies*, *47*(8), 931-936.

Karine Chalvet-Monfray

Marie Laure Delignette-Muller

2 articles très cités…

- - Bland, J. M., & Altman, D. (1986). Statistical methods for assessing agreement between two methods of clinical measurement. The lancet, 327(8476), 307-310 (***41648 citations*** *sur scholar google au 22/11/18 dont 28 fois hier*).



STATISTICAL METHODS FOR ASSESSING AGREEMENT BETWEEN TWO METHODS OF CLINICAL MEASUREMENT

J. Martin Bland, Douglas G. Altman

Department of Clinical Epidemiology and Social Medicine, St. George's Hospital Medical School, London SW17 ORE; and Division of Medical Statistics, MRC Clinical Research Centre, Northwick Park Hospital, Harrow, Middlesex

SUMMARY

In clinical measurement comparison of a new measurement technique with an established one is often needed to see whether they agree sufficiently for the new to replace the old. Such investigations are often analysed inappropriately, notably by using correlation coefficients.

- - Bland, J. M., & Altman, D. G. (2010). Statistical methods for assessing agreement between two methods of clinical measurement. International Journal of Nursing Studies, 47(8), 931-936 (*507 citations sur scholar google au 22/11/18*)

International Journal of Nursing Studies 47 (2010) 931–936

Contents lists available at ScienceDirect

International Journal of Nursing Studies

journal homepage: www.elsevier.com/ijns

ELSEVIER

Nursing Studies

Statistical methods for assessing agreement between two methods of clinical measurement☆

J. Martin Bland [a,b,*], Douglas G. Altman [a,b]

[a] Department of Clinical Epidemiology and Social Medicine, St George's Hospital Medical School, London SW17, UK
[b] Division of Medical Statistics, MRC Clinical Research Centre, Northwick Park Hospital, Harrow, Middlesex, UK

ABSTRACT

In clinical measurement comparison of a new measurement technique with an established one is often needed to see whether they agree sufficiently for the new to replace the old. Such investigations are often analysed inappropriately, notably by using correlation

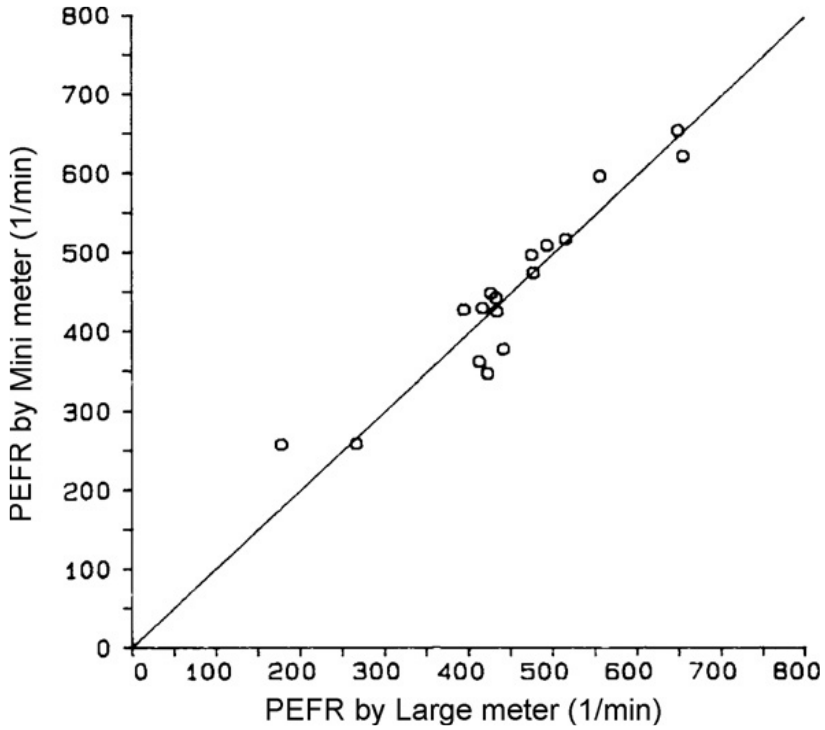
…comme quoi il faut répéter les messages

Réflexe :

- Deux variables quantitatives mesurées sur les mêmes individus…

Réflexe : **Mauvais réflexe ☹**

- Deux variables quantitatives mesurés sur les mêmes individus…



Coefficient de corrélation r

Extrait de Bland et Altman 2010, Fig 1. page 932

Pourquoi ?

- Parce que !

On va voir cela point par point.

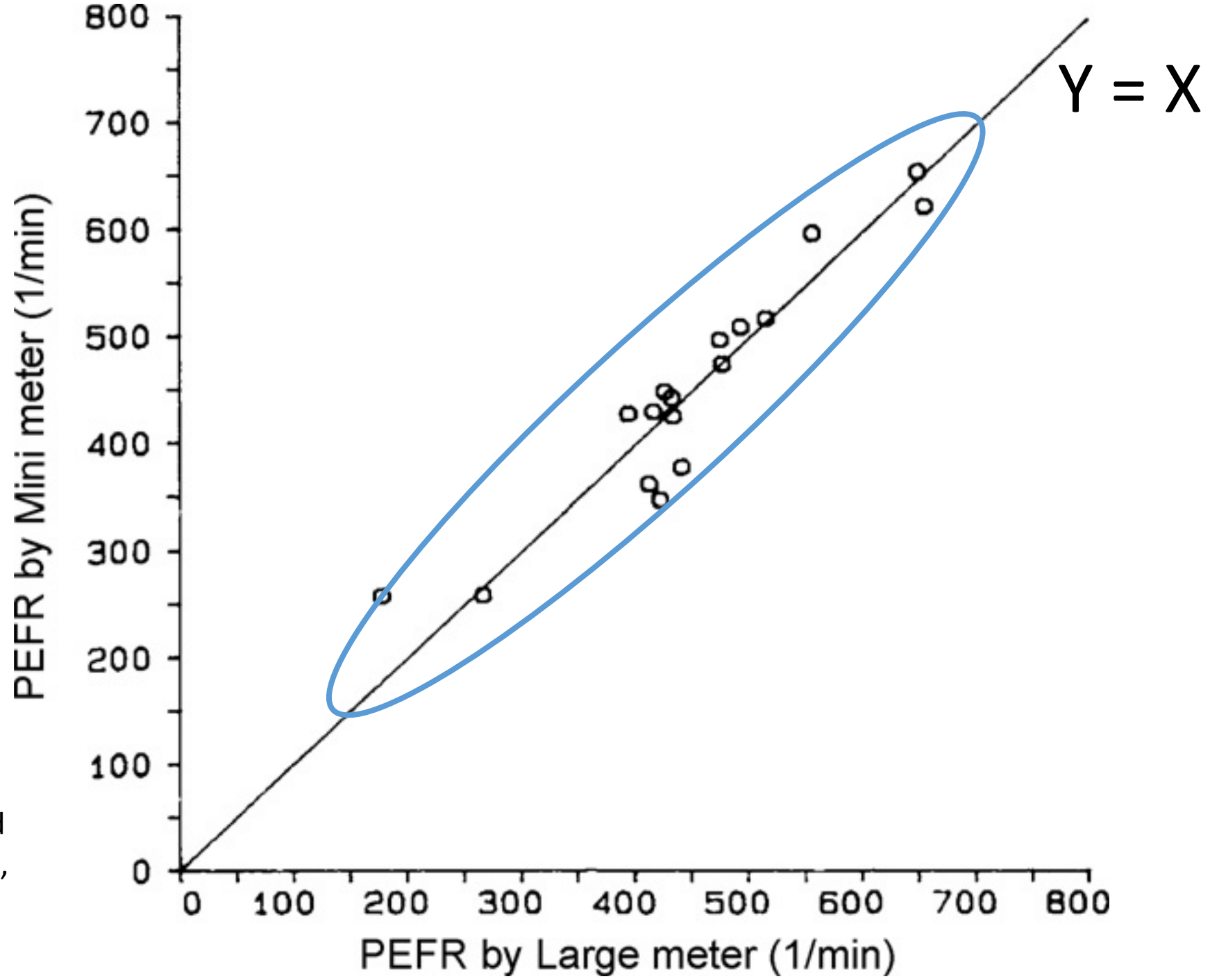
meters are related. However, this high correlation does not mean that the two methods agree:

(1) r measures the strength of a relation between two variables, not the agreement between them. We will have perfect agreement only if the points in Fig. 1 lie along the line of equality, but we will have perfect correlation if the points lie along any straight line.
(2) A change in scale of measurement does not affect the correlation, but it certainly affects the agreement. For example, we can measure subcutaneous fat by skinfold calipers. The calipers will measure two thicknesses of fat. If we were to plot calipers measurement against half-calipers measurement, in the style of Fig. 1, we should get a perfect straight line with slope 2.0. The correlation would be 1.0, but the two measurements would not agree—we could not mix fat thicknesses obtained by the two methods, since one is twice the other.
(3) Correlation depends on the range of the true quantity in the sample. If this is wide, the correlation will be greater than if it is narrow. For those subjects whose PEFR (by peak flow meter) is less than 500 l/min, r is 0.88 while for those with greater PEFRs r is 0.90. Both are less than the overall correlation of 0.94, but it would be absurd to argue that agreement is worse below 500 l/min and worse above 500 l/min than it is for everybody. Since investigators usually try to compare two methods over the whole range of values typically encountered, a high correlation is almost guaranteed.
(4) The test of significance may show that the two methods are related, but it would be amazing if two methods designed to measure the same quantity were not related. The test of significance is irrelevant to the question of agreement.
(5) Data which seem to be in poor agreement can produce quite high correlations. For example, Serfontein and Jaroszewicz (1978) compared two methods of measuring gestational age. Babies with a gestational age of 35 weeks by one method had gestations between 34 and 39.5 weeks by the other, but r was high (0.85). On the other hand, Oldham et al. (1979) compared the mini and large Wright peak flow meters and found a correlation of 0.992. They then connected the meters in series, so that both measured the same flow, and obtained a "material improvement" (0.996). If a correlation coefficient of 0.99 can be materially improved upon, we need to rethink our ideas of what a high correlation is in this context. As we show below, the high correlation of 0.94 for our own data conceals considerable lack of agreement between the two instruments.

A propos de l'utilisation fréquente du coefficient de corrélation (1) - p. 932

"r measures the strength of a relation between two variables, not the agreement between them. We will have perfect agreement only if the points in Fig. 1 lie along the line of equality, but we will have perfect correlation if the points lie along any straight line."

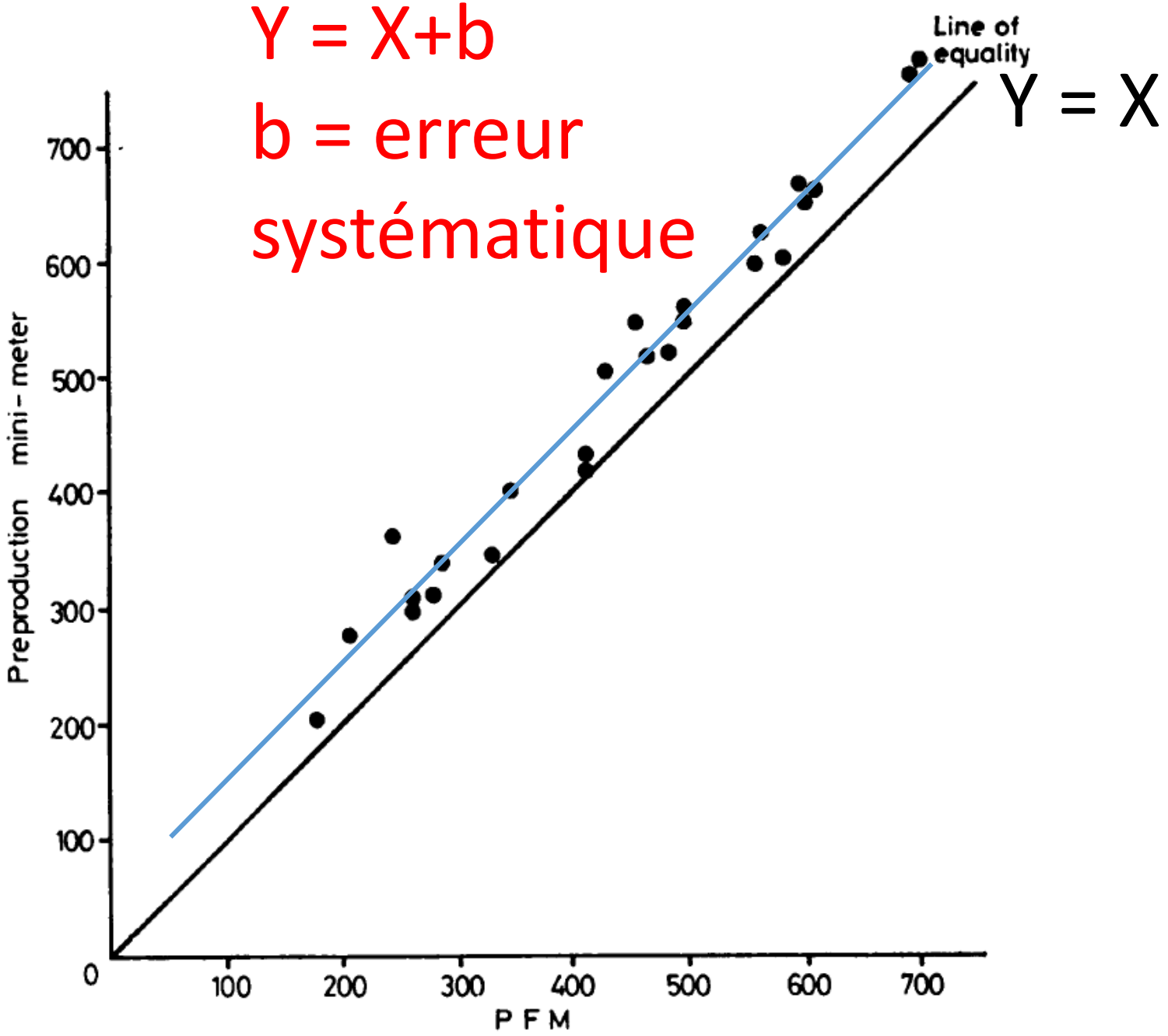
« r mesure la force d'une relation entre deux variables, pas l'accord entre elles. L'accord ne sera parfait que si les points de la Fig. 1 se situent le long de la ligne d'égalité, mais nous aurons **une corrélation parfaite si les points se trouvent le long de n'importe quelle ligne droite** [sauf horizontale]). »



Extrait de Bland et Altman 2010, Fig 1. page 932

Y = X+b

b = erreur systématique

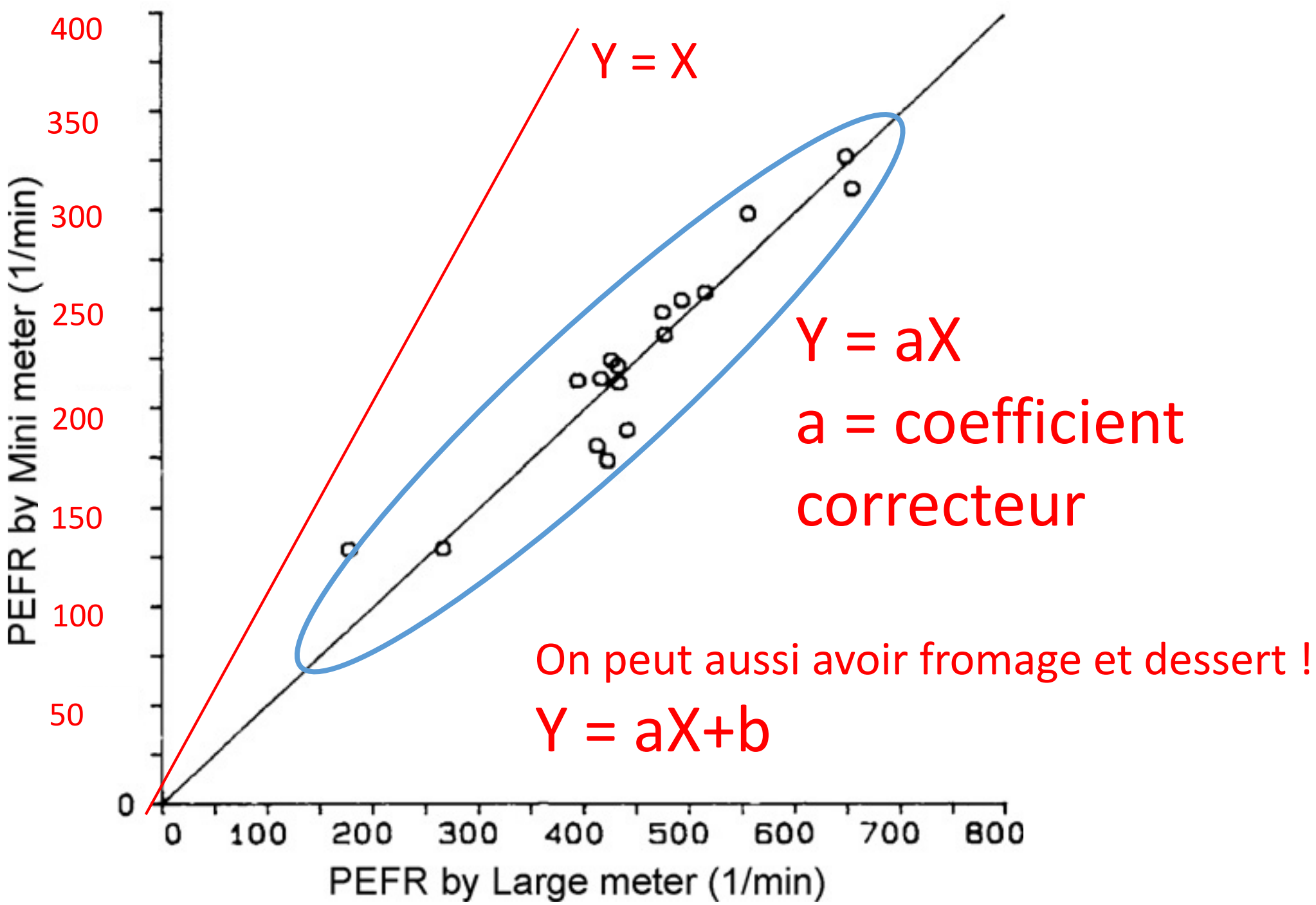


D'après
Oldham 1979
Fig 1

A propos de l’utilisation fréquente du coefficient de corrélation (2) - p. 932

“A change in scale of measurement does not affect the correlation, but it certainly affects the agreement.”

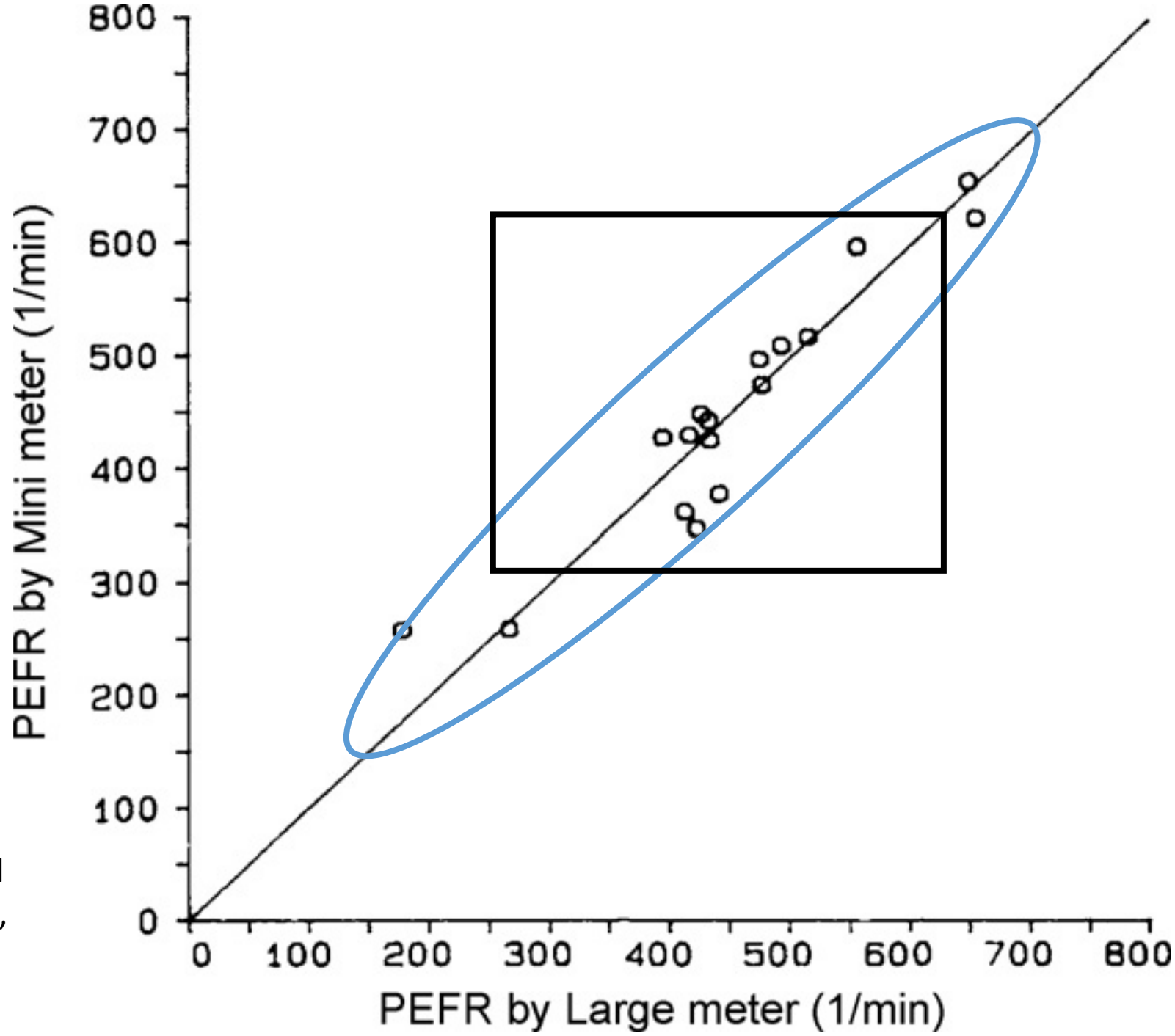
« **Un changement d'échelle de mesure n'affecte pas la corrélation, mais affecte certainement l'accord**. »


Y = X
Y = aX
a = coefficient
correcteur
On peut aussi avoir fromage et dessert !
Y = aX+b
PEFR by Mini meter (1/min)
PEFR by Large meter (1/min)
0
50
100
150
200
250
300
350
400
0
100
200
300
400
500
600
700
800

A propos de l'utilisation fréquente du coefficient de corrélation (3) - p. 932

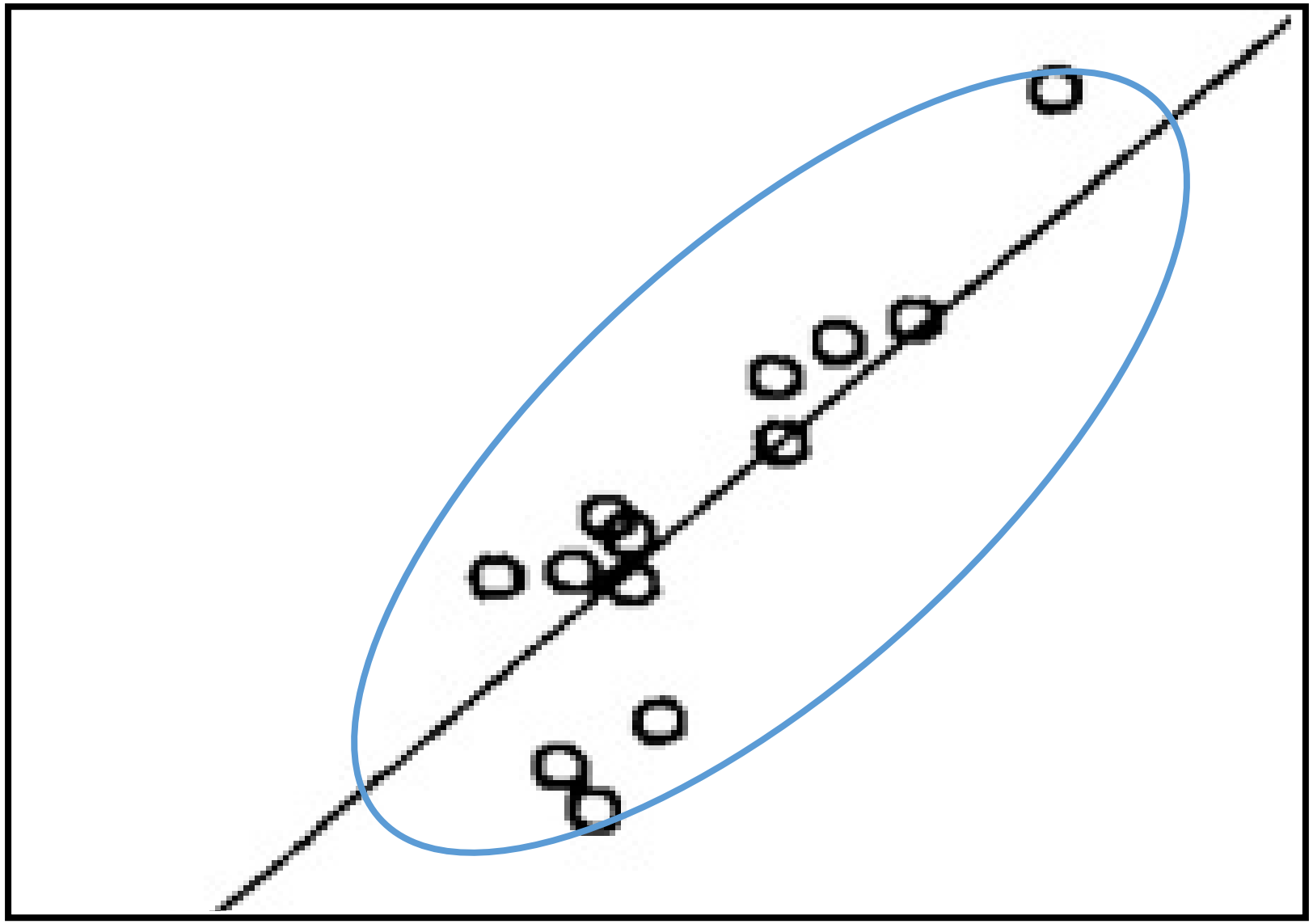
"Correlation depends on the range of the true quantity in the sample. If this is wide, the correlation will be greater than if it is narrow... Since investigators usually try to compare two methods over the whole range of values typically encountered, a high correlation is almost guaranteed."

« La corrélation dépend de la gamme de variation de la quantité réelle dans l'échantillon. Si celle-ci est large, la corrélation sera plus grande que si elle est étroite ... Etant donné que les enquêteurs tentent généralement de comparer deux méthodes **sur toute la gamme de variation des valeurs généralement rencontrées, une corrélation élevée est presque garantie.** »



Extrait de Bland et Altman 2010, Fig 1. page 932

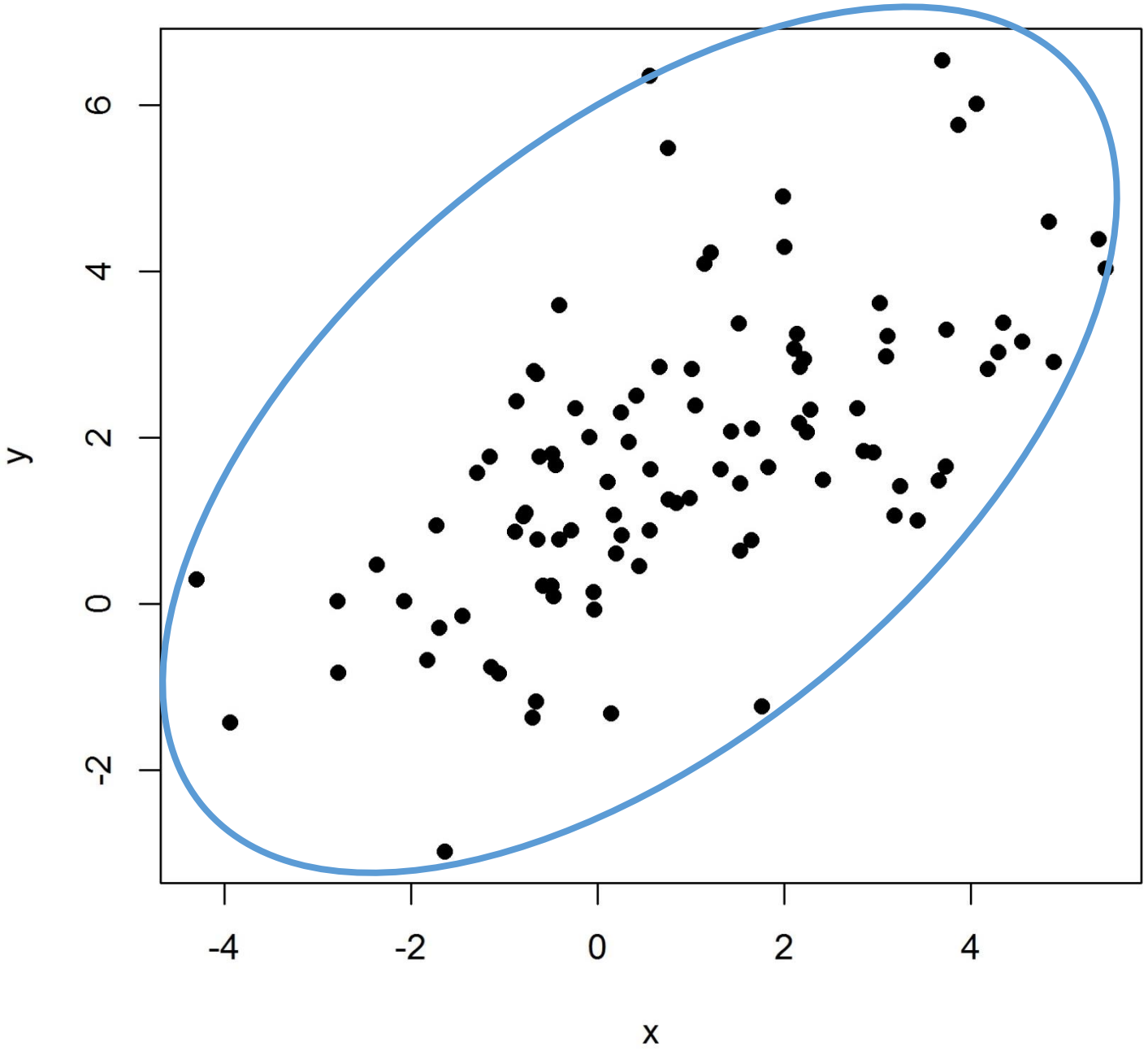
Si on zoome la corrélation sera moins bonne ⇔ r plus faible



A propos de l'utilisation fréquente du coefficient de corrélation (4) - p. 932

"The test of significance may show that the two methods are related, but it would be amazing if two methods designed to measure the same quantity were not related. The test of significance is irrelevant to the question of agreement."

« Le test de signification peut montrer que les deux méthodes sont liées, mais il serait étonnant que deux méthodes conçues pour mesurer la même quantité ne soient pas liées. **Le critère de signification n'a aucun intérêt sur la question de l'accord**. »



r = 0.61

Test de corrélation

p = 7.6 10^{-12}

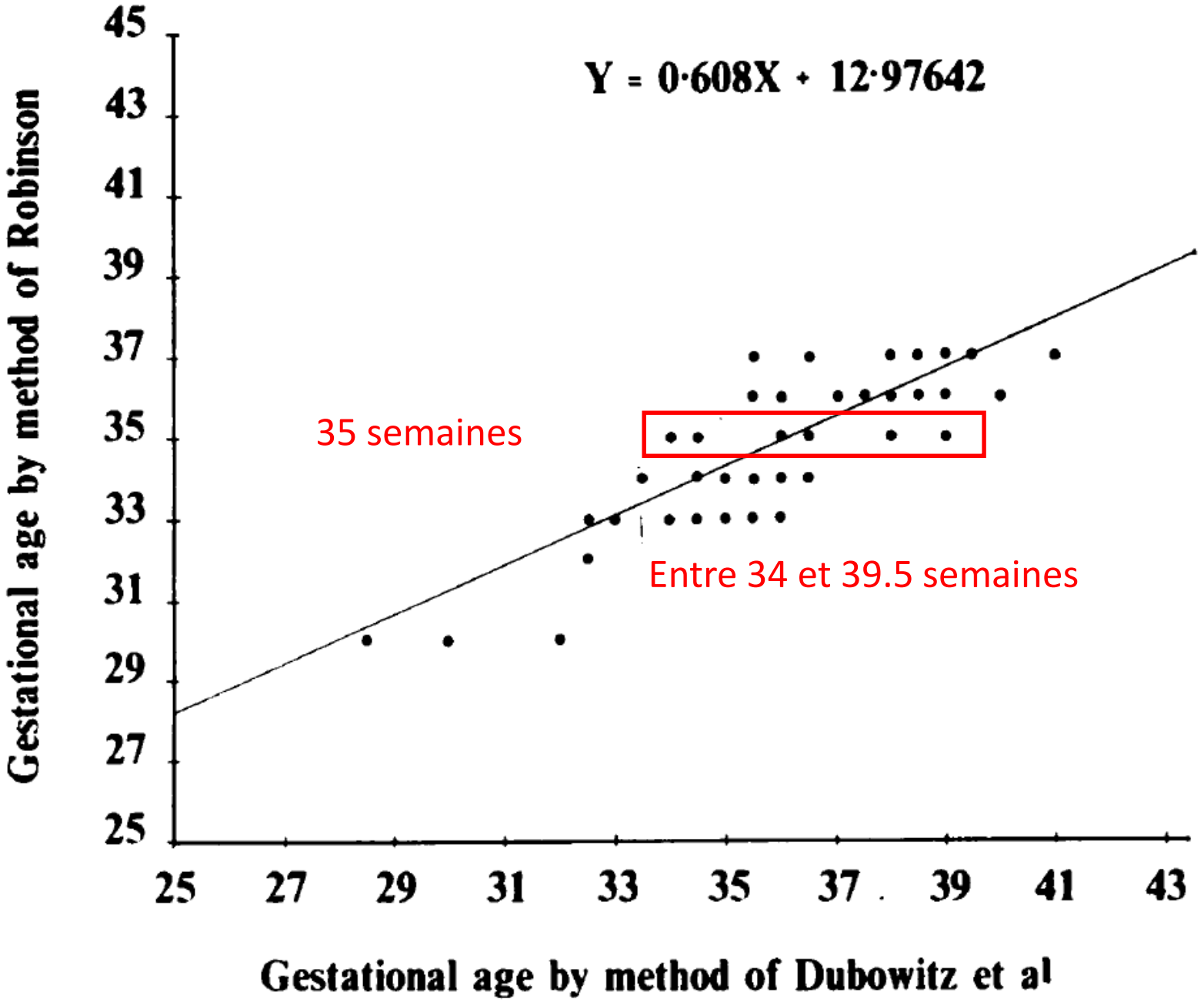
Corrélation très significative… mais

un tel résultat vous satisferait-il ?

A propos de l'utilisation fréquente du coefficient de corrélation (5a) - p. 932

"Data which seem to be in poor agreement can produce quite high correlations. For example, Serfontein and Jaroszewicz (1978) compared two methods of measuring gestational age. Babies with a gestational age of 35 weeks by one method had gestations between 34 and 39.5 weeks by the other, but r was high (0.85)."

« **Les données qui semblent être en mauvais accord peuvent produire des corrélations assez élevées.** Par exemple, Serfontein et Jaroszewicz (1978) ont comparé deux méthodes de mesure de l'âge gestationnel. Les bébés dont l'âge gestationnel était de 35 semaines selon l'une des méthodes avaient des gestations comprises entre 34 et 39,5 semaines avec l'autre méthode, mais r était élevé (0,85). »



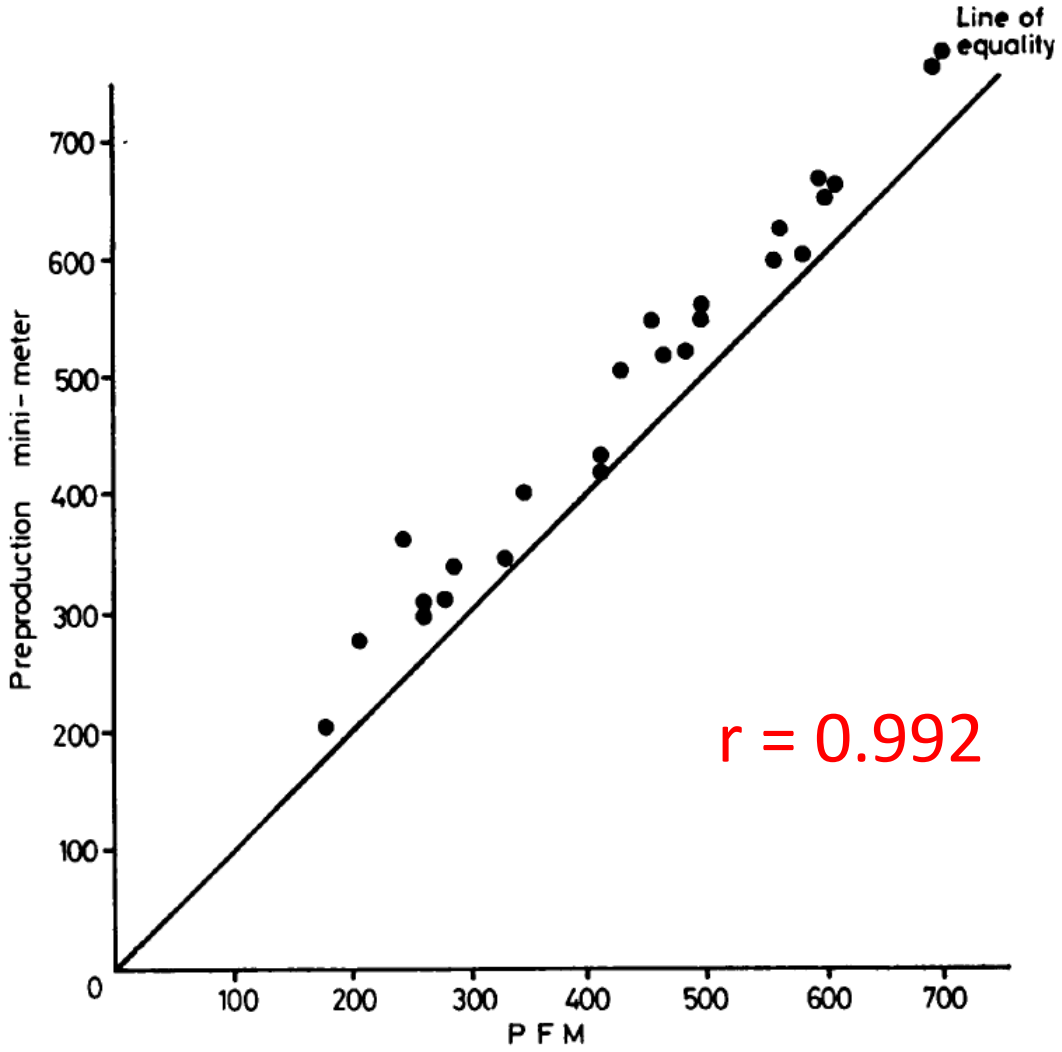
r = 0.85

Extrait de Serfontein and Jaroszewicz 1978

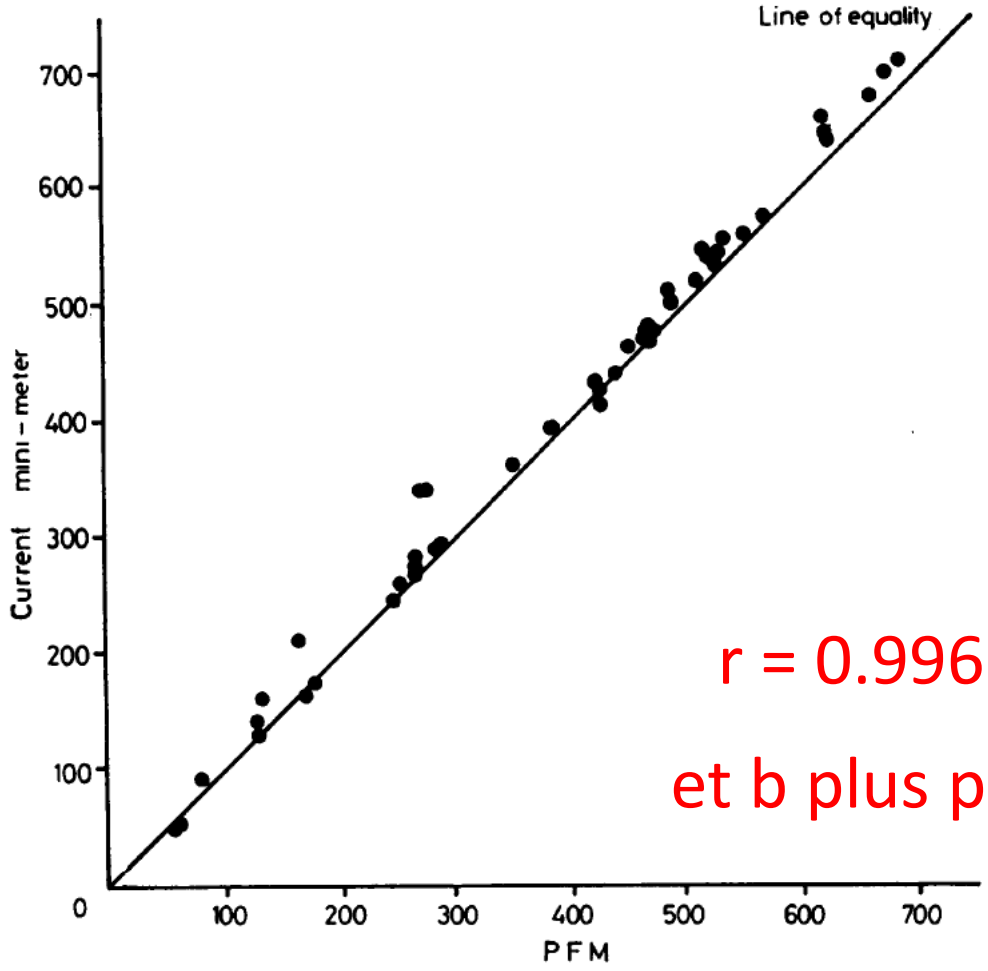
A propos de l'utilisation fréquente du coefficient de corrélation (5b) - p. 932

"On the other hand, Oldham et al. (1979) compared the mini and large Wright peak flow meters and found a correlation of 0.992. They then connected the meters in series, so that both measured the same flow, and obtained a ''material improvement'' (0.996). If a correlation coefficient of 0.99 can be materially improved upon, we need to rethink our ideas of what a high correlation is in this context."

« D'autre part, Oldham et al. (1979) ont comparé les débitmètres de pointe mini et grand de Wright et ont trouvé une corrélation de 0,992. Ils ont ensuite connecté les compteurs en série, de sorte que les deux mesurent le même débit et obtiennent une «amélioration matérielle» (0,996). **Si un coefficient de corrélation de 0,99 peut être sensiblement amélioré, nous devons repenser nos idées sur ce qu'est une corrélation élevée dans ce contexte.** »



r = 0.992



r = 0.996

et b plus petit

Par une modification technique

Extrait de
Oldham 1979
Fig 1 et 4

Que faut-il faire alors ? - p. 932

“It is most unlikely that different methods will agree exactly, by giving the identical result for all individuals. We want to know by how much the new method is likely to differ from the old; if this is not enough to cause problems in clinical interpretation we can replace the old method by the new or use the two interchangeably.”

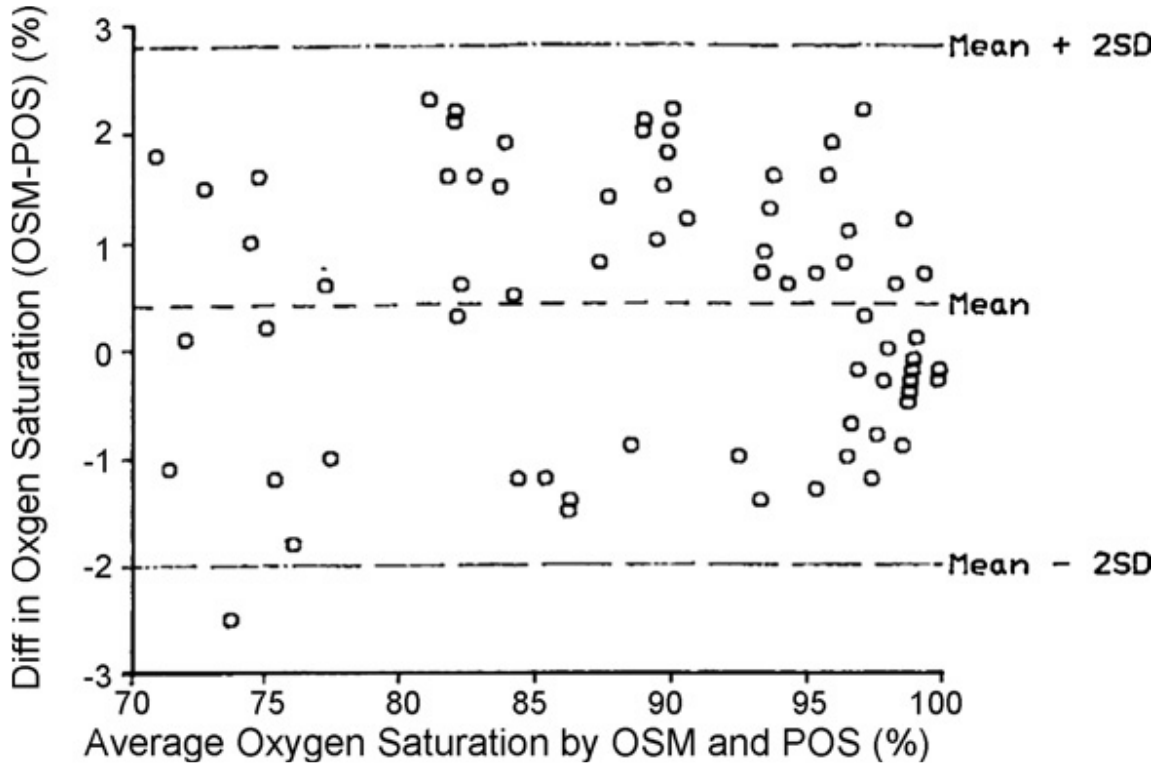
« Il est très peu probable que différentes méthodes s'accordent exactement, en donnant le même résultat pour tous les individus. Nous voulons savoir dans quelle mesure la nouvelle méthode est susceptible de différer de l’ancienne; **si cela n’est pas suffisant pour poser des problèmes d’interprétation clinique, nous pouvons alors remplacer l’ancienne méthode par la nouvelle ou utiliser les deux de manière interchangeable**. »

Réflexe :

- Deux variables quantitatives mesurées sur les mêmes individus…

Réflexe : **Bon réflexe ☺**

- Deux variables quantitatives mesurés sur les mêmes individus…



Extrait de Bland et Altman 2010, Fig 3 - p. 933

Graphe des différences en fonction de la moyenne des deux méthodes - p. 933

“The first step is to examine the data. A simple plot of the results of one method against those of the other (Fig. 1) though without a regression line is a useful start but usually all the data points will be clustered near the line and it will be difficult to assess between-method differences. A plot of the difference between the methods against their mean may be more informative.”

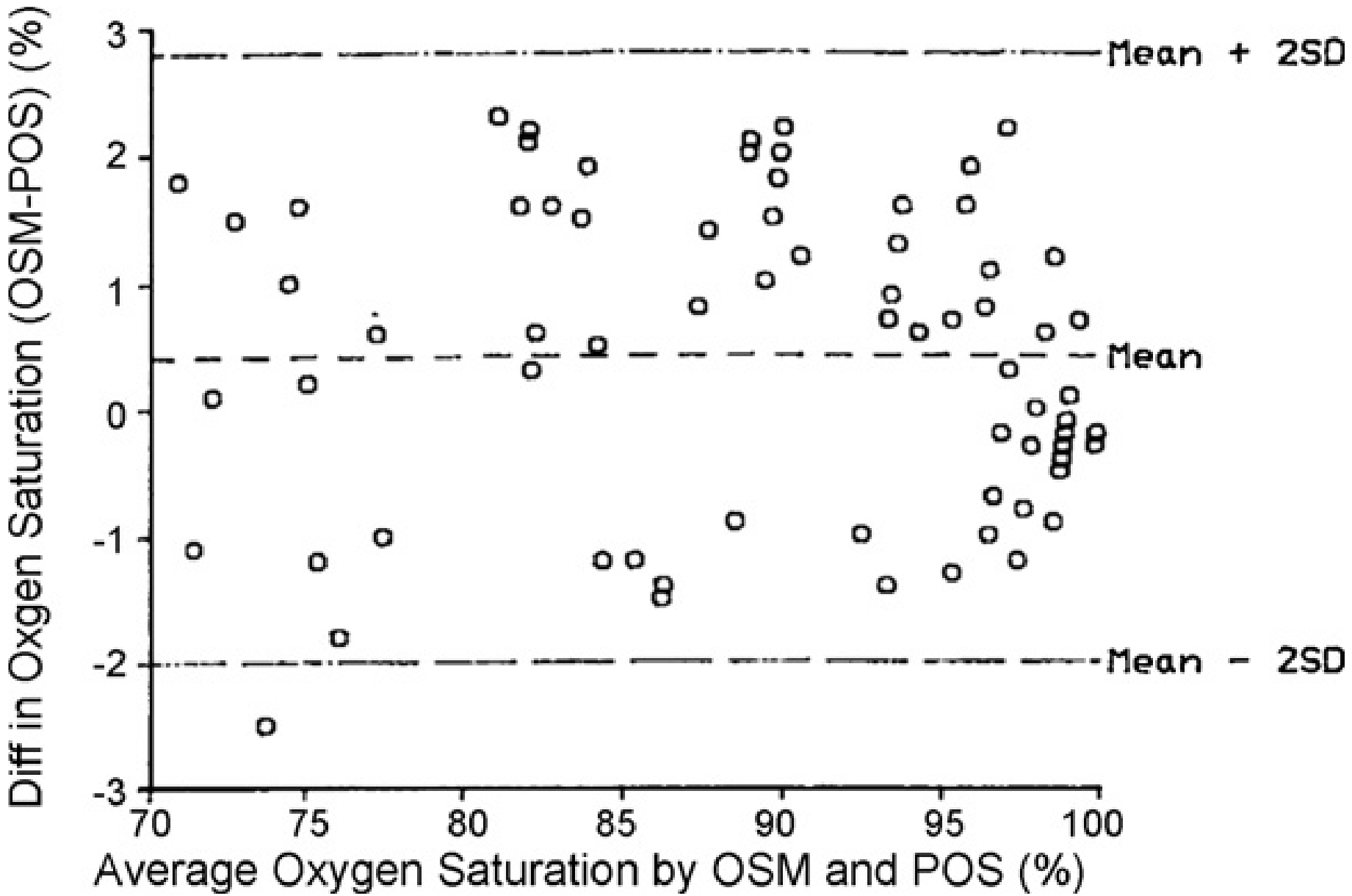
« La première étape consiste à examiner les données. Un simple graphique des résultats d'une méthode par rapport à ceux de l'autre (Fig. 1), bien que l'absence d'une ligne de régression soit un début utile, mais généralement tous les points de données sont regroupés près de la ligne et il sera difficile d'évaluer une différence entre les méthodes. **Un graphique de la différence entre les méthodes en fonction de leur moyenne peut être plus informatif**. »

Graphe des différences en fonction de la moyenne des deux méthodes – p. 933

"Provided differences within $\bar{d} \pm 2\hat{\sigma}_d$ *would not be clinically important we could use the two measurement methods interchangeably. We shall refer to these as the 'limits of agreement'. "*

« À condition que les différences comprises entre $\bar{d} \pm 2\hat{\sigma}_d$ **ne soient pas cliniquement importantes**, nous pourrions utiliser les deux méthodes de mesure de manière interchangeable. Nous qualifierons **[ces limites] de 'limites d'accord'**. »

△ Intervalle de fluctuation des différences calculable ainsi si la distribution des différences est proche d'une loi normale

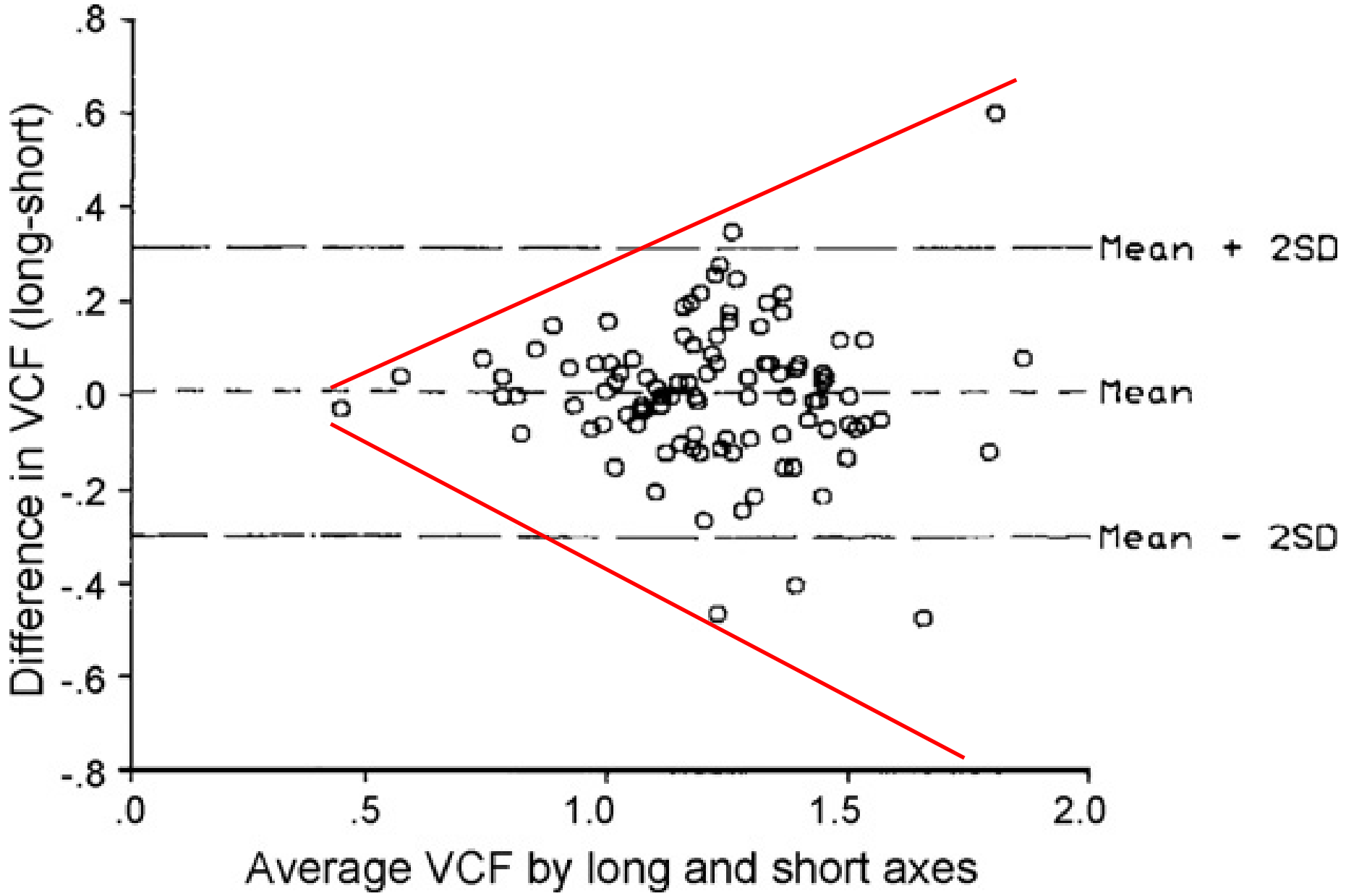


Extrait de Bland et Altman 2010, Fig 3 p.933

Transformation logarithmique nécessaire si la différence est proportionnelle à la moyenne – p.934

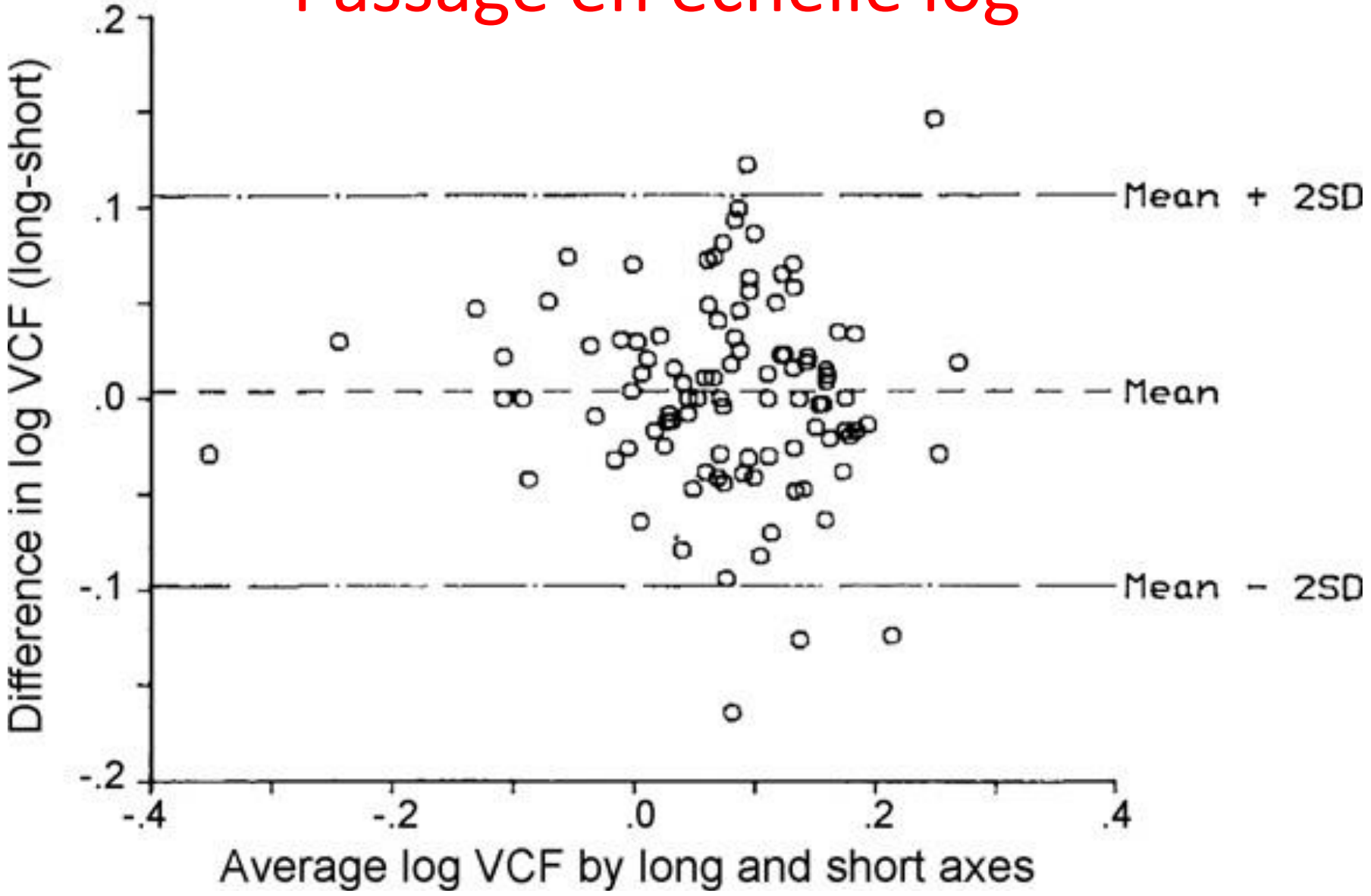
“If the differences are proportional to the mean, a logarithmic transformation should yield a picture more like that of Figs. 2 and 4, and we can then apply the analysis described above to the log transformed data.”

« Si les **différences sont proportionnelles à la moyenne, une transformation logarithmique** devrait donner une image plus proche de celle des fig. 2 et 4, et nous pouvons ensuite appliquer l'analyse décrite ci-dessus aux données transformées en log. »



Extrait de Bland et Altman 2010, Fig 4. p. 934

Passage en échelle log



Extrait de Bland et Altman 2010, Fig 5. p. 934

Vérification de la normalité de la distribution des différences – complément

Les méthodes graphiques que vous connaissez pour juger de la normalité d'une distribution (visualisation de la distribution sous forme d'histogramme si on a beaucoup de données, ou de diagramme en boîte **+ diagramme Quantile-Quantile**) pourrons bien entendu être utilisées en complément de la figure de base proposée par Bland et Altman.

Maintenant à vous de jouer sur un exemple …