Evaluation de tâches d'estimation sans gold standard

Irène Buvat
Imagerie et Modélisation en Neurobiologie et Cancérologie
UMR 8165 CNRS - Universités Paris 7 et Paris 11

buvat@imnc.in2p3.fr

5 mars 2009

Introduction

En présence de gold standard :
- biais
- variabilité
- combinaison des 2 critères (RMSE, etc…)
- corrélation

En l'absence de gold standard ?

Plan

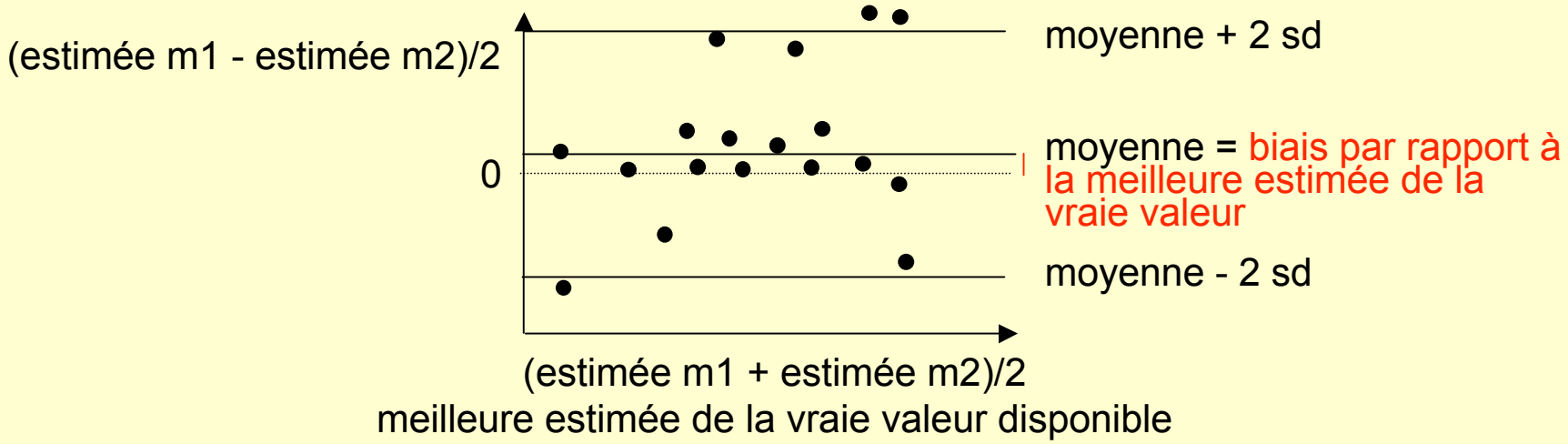
- Approche Bland-Altman

- Régression dans gold standard

- Evaluation simultanée du gold standard et des performances de méthodes de segmentation d'images (STAPLE)

Approche de Bland-Altman (1)

Évaluation de l'**accord** entre deux méthodes



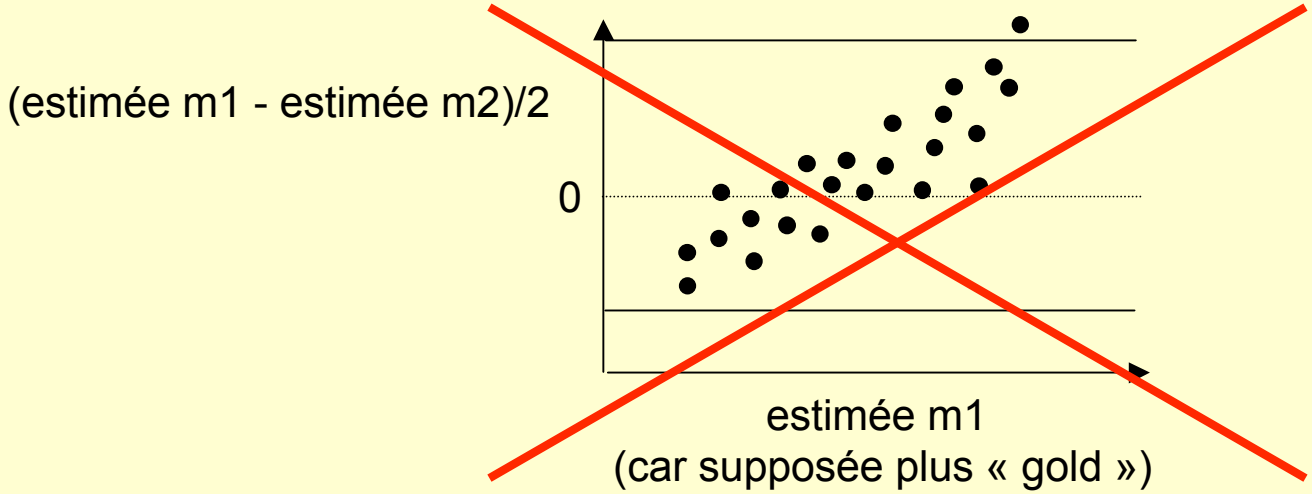
La plupart des différences sont comprises dans l'intervalle [moyenne - 2 sd ; moyenne + 2 sd]

L'étendue de cet intervalle doit permettre de conclure à l'interchangeabilité des méthodes ou non, mais PAS au fait que l'une est moins biaisée que l'autre !

Bland and Altman. Lancet: 307-10, 1986.

Approche de Bland-Altman (2)

Permet de détecter des différences systématiques entre les méthodes

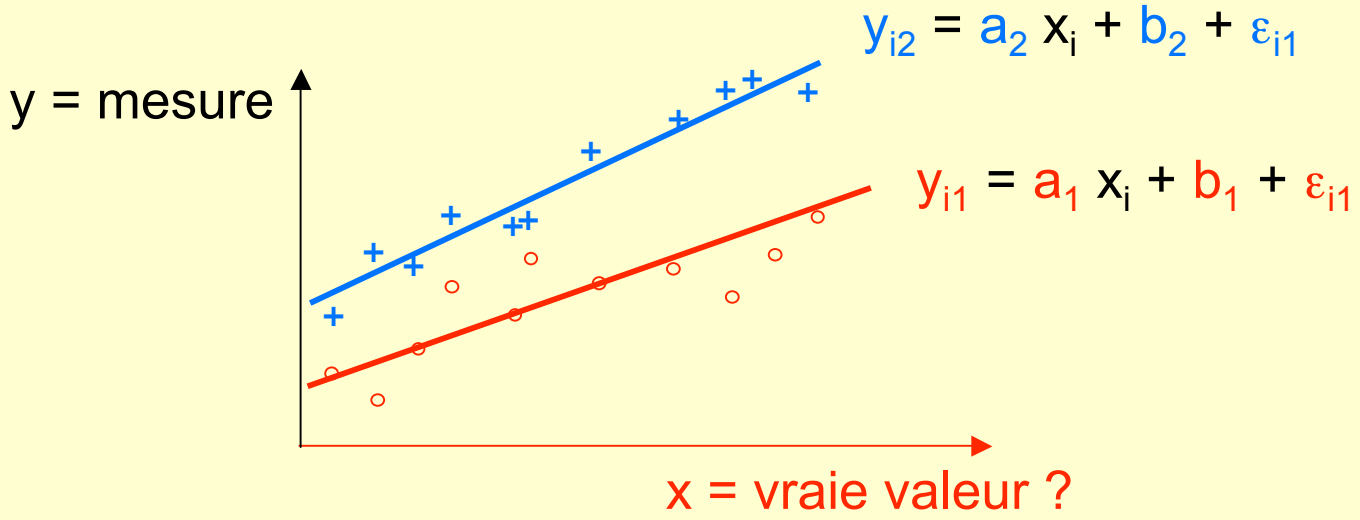


Bland and Altman. Lancet, 346: 1085-7, 1995.

EWAGS (1)

Hypothèses:

1) Au moins 2 méthodes m d'estimation du même paramètre x_i



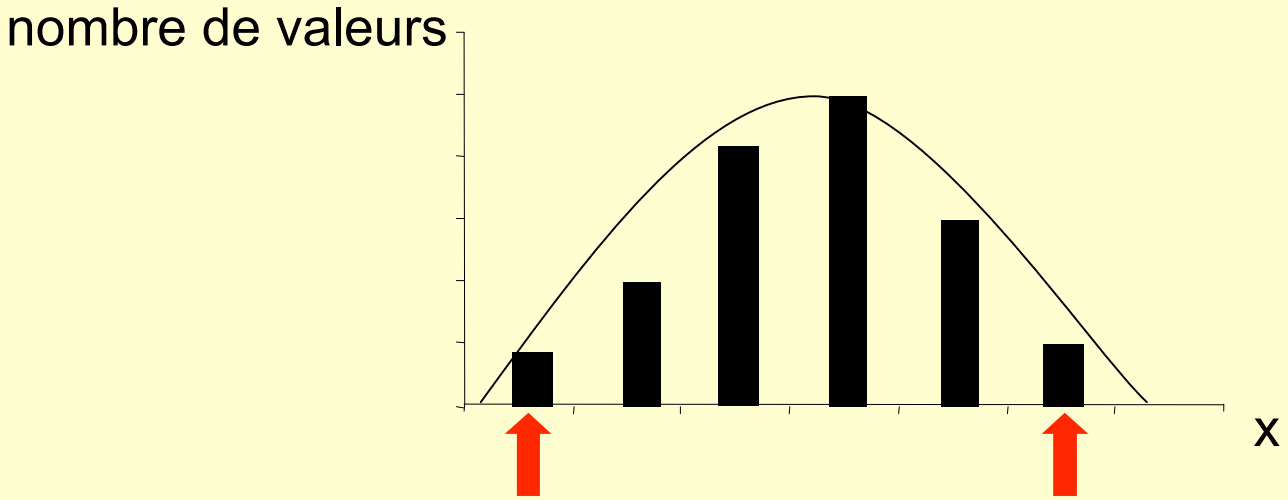
2) $y_{im} = a_m x_i + b_m + \varepsilon_{mi}$ pour m = 1, 2, … et $\varepsilon_{mi} \sim \mathcal{N}(0,\sigma_m)$

Hoppin, Kupinski, Kastis, Clarkson, Barrett. IEEE Trans Med Imaging 21: 441-449, 2002
Kupinski, Hoppin, Clarkson, Barrett, Kastis. Acad Radiol 9: 290-297, 2002

EWAGS (2)

Hypothèses :

3) x_i suit une loi beta ou une distribution normale tronquée θ définie par 2 paramètres inconnus π_1 et π_2.
Les valeurs min et max de la distribution sont approximativement connues



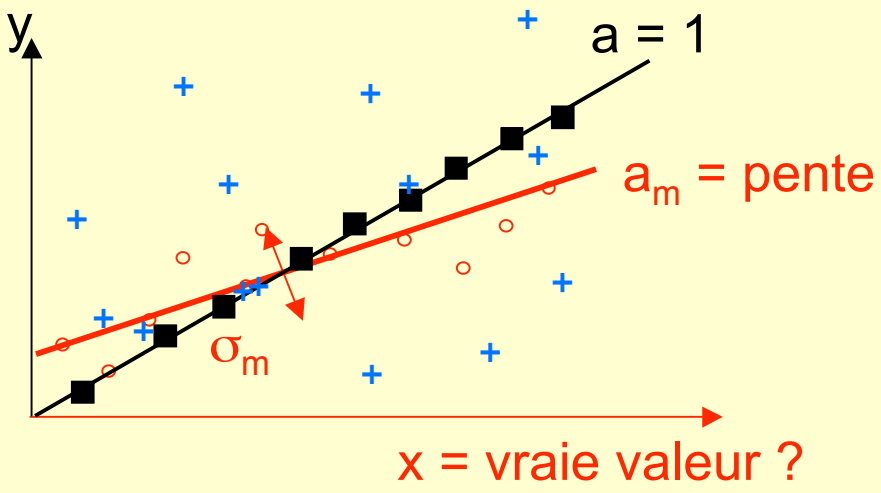
EWAGS (3)

Méthodes :

1) Maximisation de la log-vraisemblance des paramètres du modèle

$$\mathcal{L}(\{a_m, b_m, \sigma_m\}_{\text{pour tt m}} \mid \{y_{im}\})$$

2) Calcul de σ_m/a_m comme figure de mérite



Estimée idéale :
- pas de dispersion : $\sigma_m \sim 0$
- pente = a_m =1
- σ_m/a_m tend vers 0

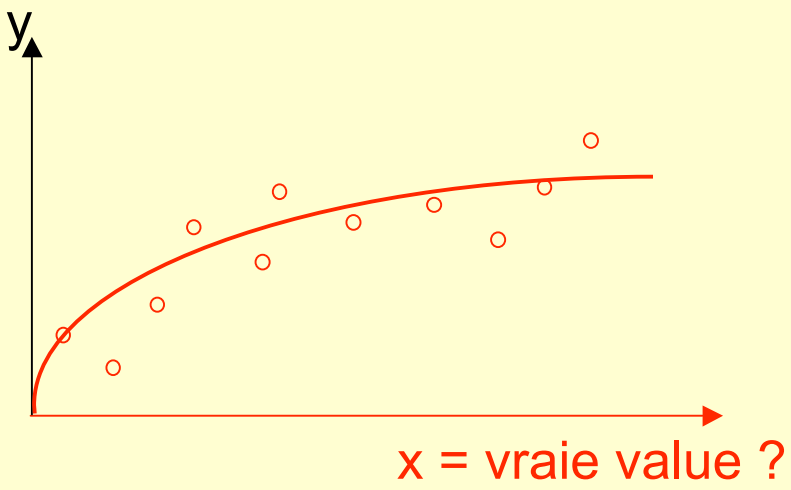
Estimée sans intérêt :
- grande dispersion : $\sigma_m >> 1$
- pente = $a_m \sim 0$
- σ_m/a_m tend vers ∞

Généralisation de EWAGS : GEWAGS (1)

Hypothèses :

1) Au moins 2 méthodes m d'estimation du même paramètre x_i

2) $y_{im} = a_m x_i^2 + b_m x_i + c_m + \varepsilon_{mi}$ pour m = 1, 2, … et $\varepsilon_{mi} \sim \mathcal{N}(0,\sigma_m)$

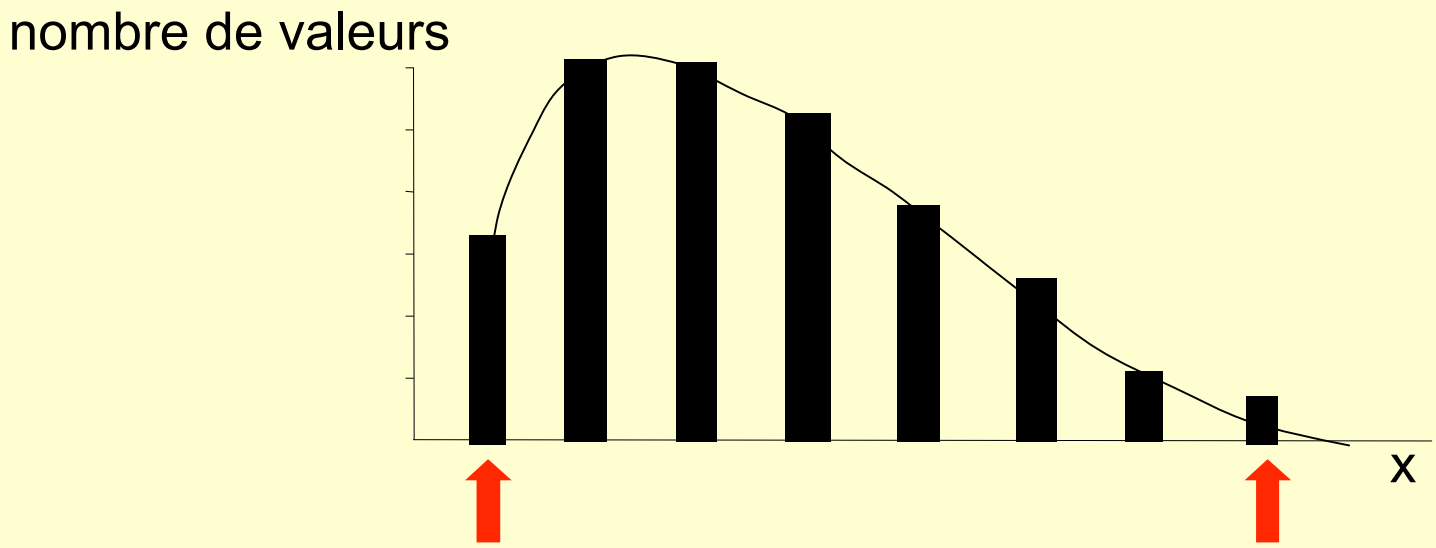


Buvat et al, J Nucl Med 48: 44P, 2007

Généralisation de EWAGS : GEWAGS (2)

Hypothèses :

3) x_i suit une distribution beta θ définie par 2 paramètres inconnus π_1 et π_2
Les valeurs min et max de la distribution sont approximativement connues



Pas de choix à faire entre loi beta et loi gaussienne
Rend compte des distributions asymétriques

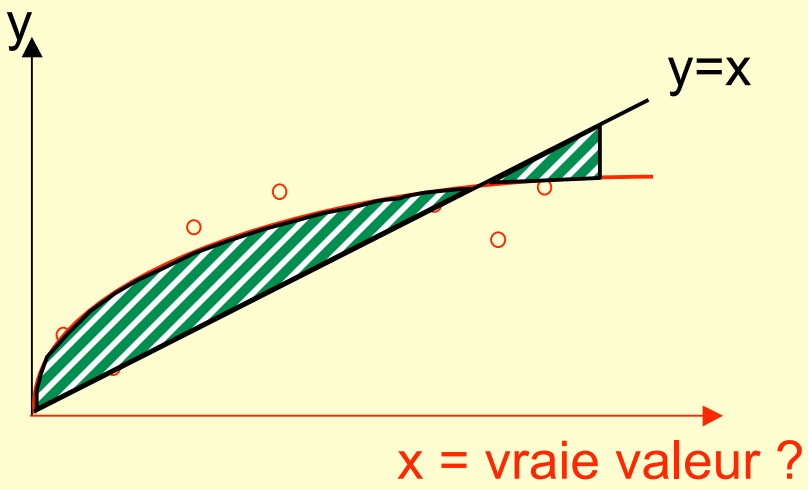
Généralisation de EWAGS : GEWAGS (3)

Méthode :

1) Maximisation de la log-vraisemblance des paramètres du modèle

$$\mathcal{L}(\{a_m, b_m, c_m, \sigma_m\} \mid \{y_{im}\})$$

2) Calcul de $sMSE_m = \Sigma_{i=1,X}(y_{im}-x_i)^2/X$ comme figure de mérite



3) Calcul d'intervalles de confiance autour de $sMSE_m$ au moyen d'une approche bootstrap non paramétrique

Discussion

- Plus on dispose de méthodes d'estimation répondant au modèle, meilleur sera l'ajustement, mais 2 méthodes sont suffisantes

- Au moins 25 cas sont nécessaires

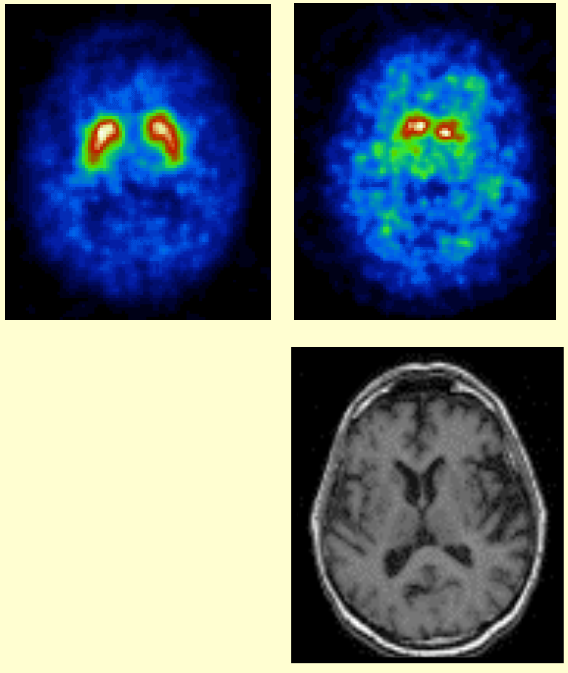
- La méthode est robuste même lorsque l'hypothèse sur la distribution des x_i est approximative

Hoppin, Kupinski, Kastis, Clarkson, Barrett. IEEE Trans Med Imaging 21: 441-449, 2002
Kupinski, Hoppin, Clarkson, Barrett, Kastis. Acad Radiol 9: 290-297, 2002

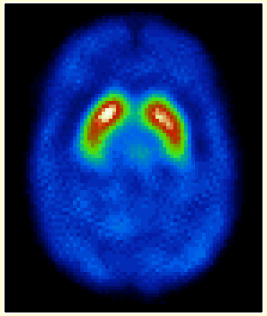
Exemple d'application 1 - Données

Images 123I-FP-CIT SPECT corrigées de l'atténuation et de la diffusion

23 patients
(13 MA + 10 DCL)



23 patients simulés*



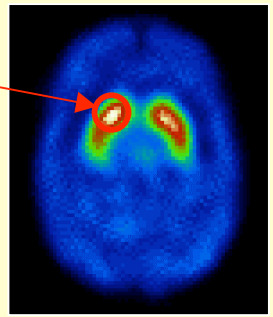
Cartographie de densité des tissus déduites des IRM des patients

Cartographie d'activité déduites des images SPECT de patients

**Soret, Koulibaly, Darcourt, Buvat. Eur J Nucl Med Mol Imaging 33:1062-1072, 2006*

Exemple d'application 1 - Méthodes comparées

Mesure du potentiel de liaison (PL) =
(activité striata - activité fond)/ activité fond



ds les 2 noyaux caudé et les 2 putamens

2 méthodes à comparer :

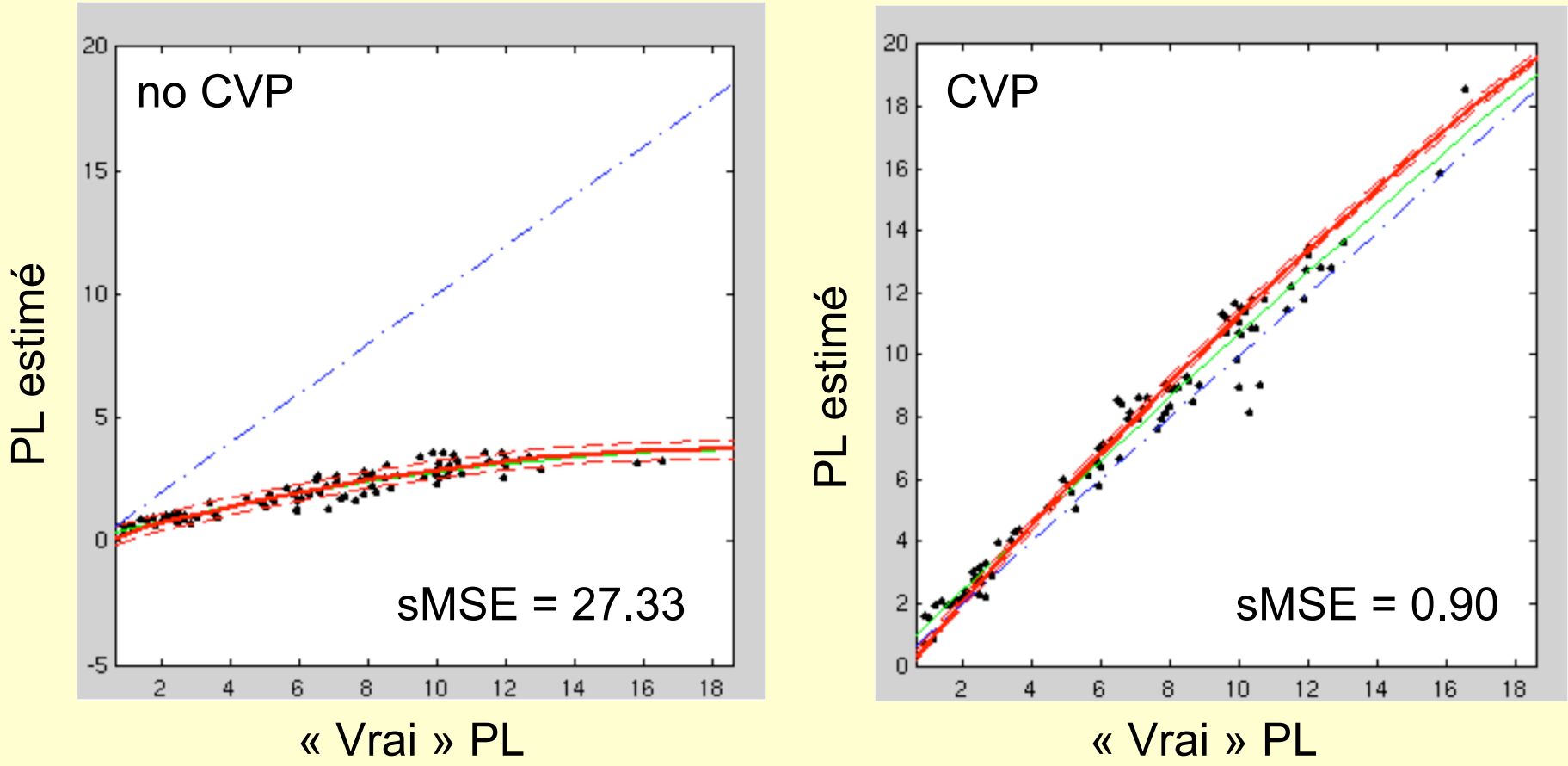
mesure sans correction de volume partiel (no CVP)
mesure avec correction de volume partiel * (CVP)

Mise en œuvre de EWAGS et GEWAGS pour comparer les 2 méthodes

**Soret, Koulibaly, Darcourt, Hapdey, Buvat. J Nucl Med 44: 1184-1193, 2003*

Exemple d'application 1 - Résultats de GEWAGS (1)

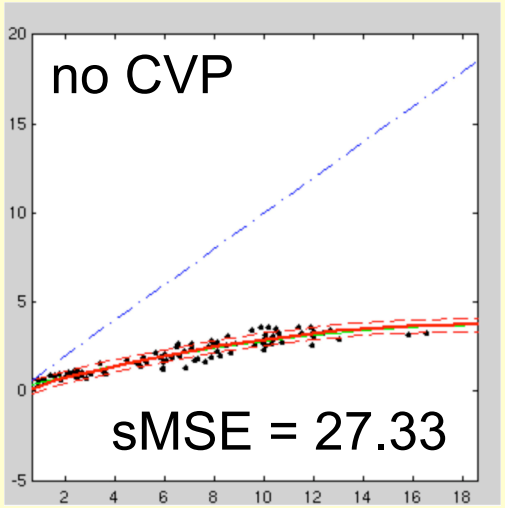
Patients simulés

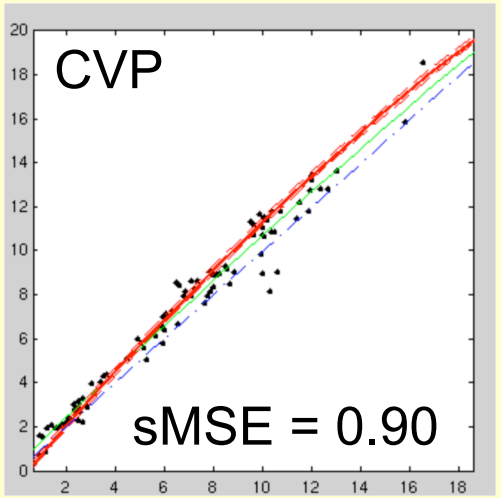


Soret, Alaoui, Koulibaly, Darcourt, Buvat. Nucl Instrum Meth Phys Res A 571: 173-176, 2007

Exemple d'application 1 - Résultats de GEWAGS (2)

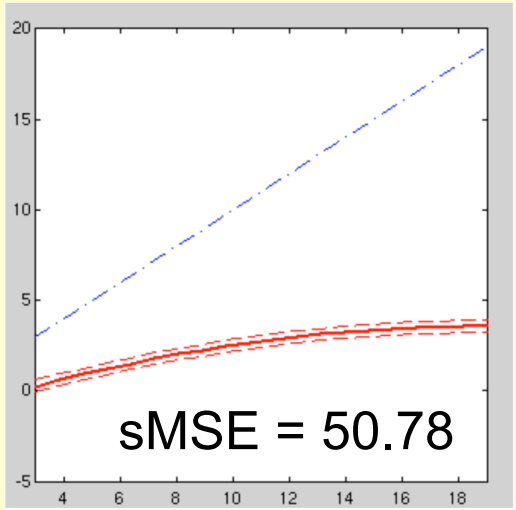
Patients simulés

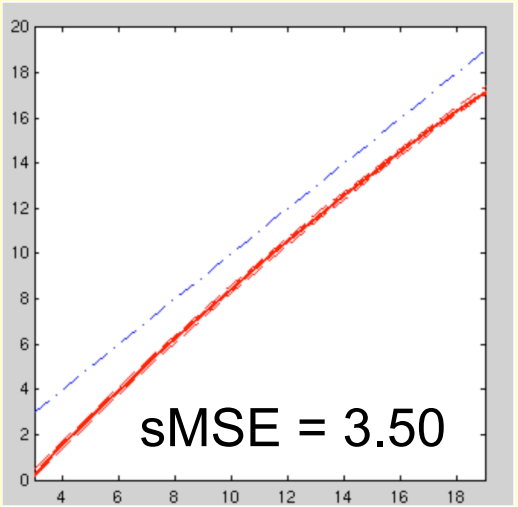




p<0.01

Patients réels

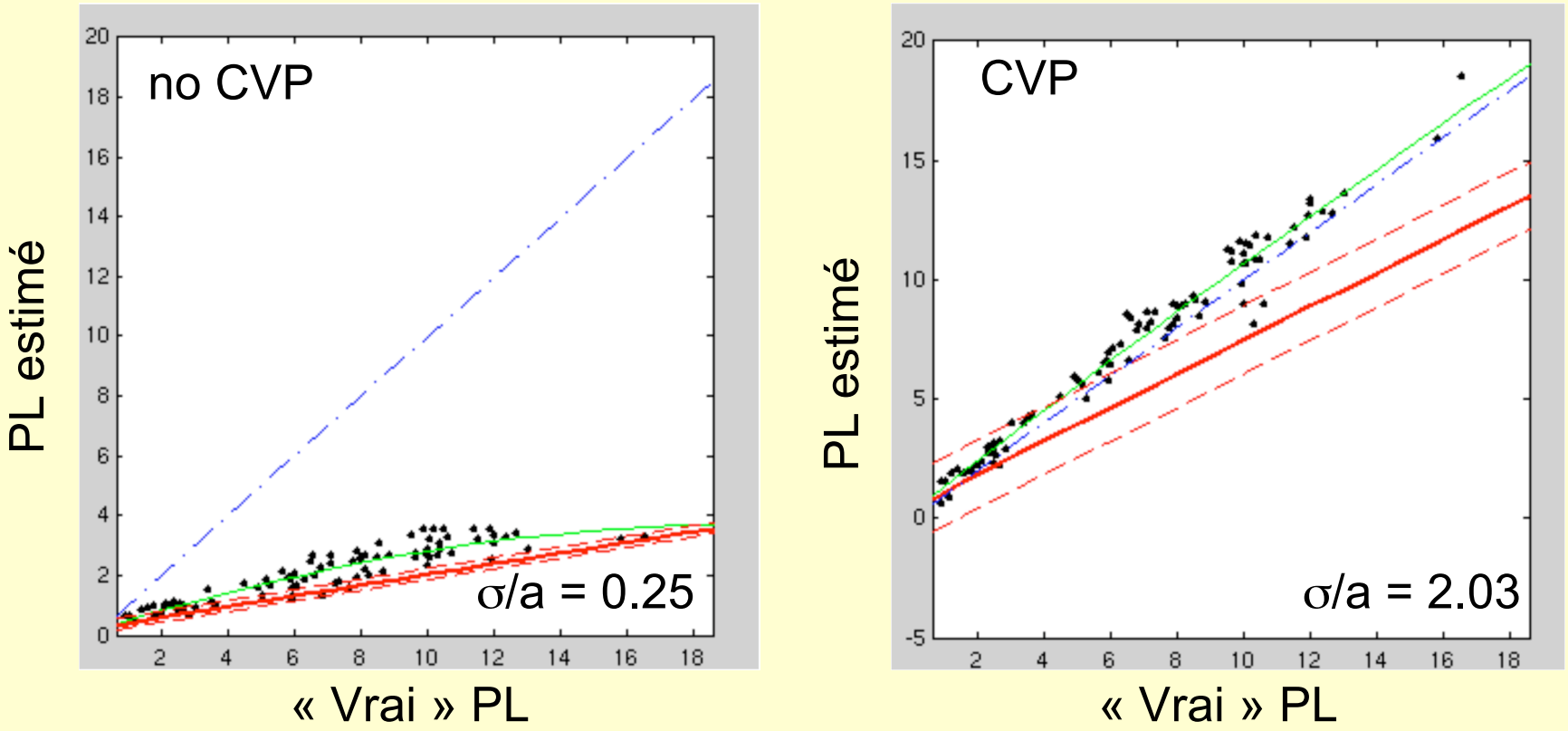




p<0.01

Exemple d'application 1 - Limites de EWAGS

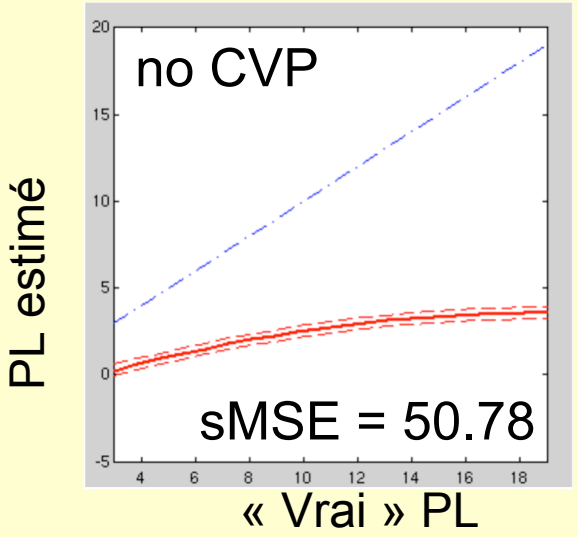
Patients simulés

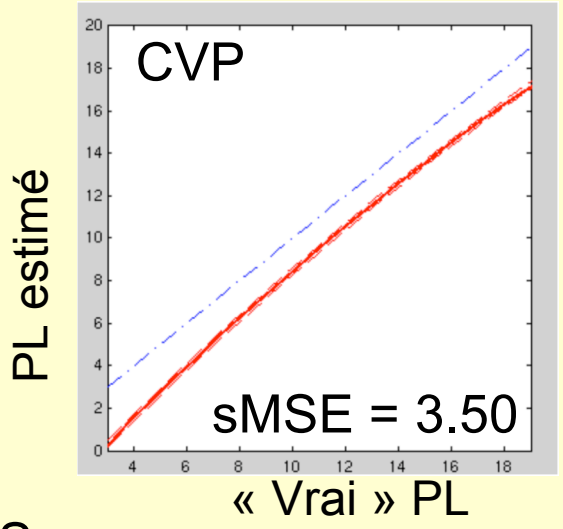


Exemple d'application 1 - Limites de EWAGS

Patients réels

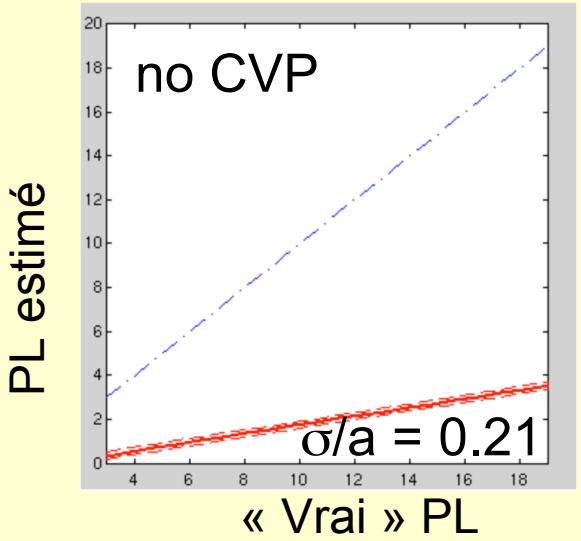
GEWAGS

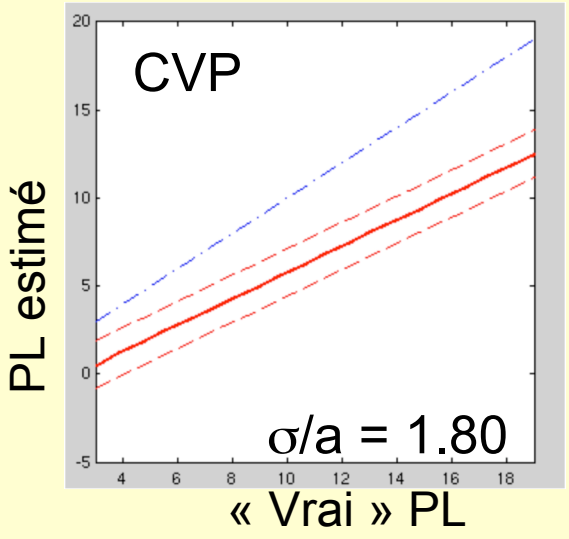




p<0.01

EWAGS





Robustesse de GEWAGS

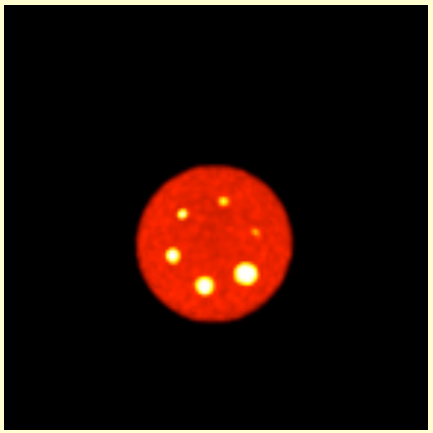
	GEWAGS sMSE		EWAGS σ/a	
	no CVP	CVP	no CVP	CVP
MA simulés	66.6	2.6*	0.22	2.80
DCL simulés	11.4	2.2*	0.17	0.08
MA+DCL simulés	27.3	0.9*	0.25	2.03
MA réels	35.4	2.9*	0.16	0.06
DCL réels	8.46	1.6*	0.17	0.07
MA+DCL réels	50.8	3.5*	0.21	1.80

* $p<0.01$

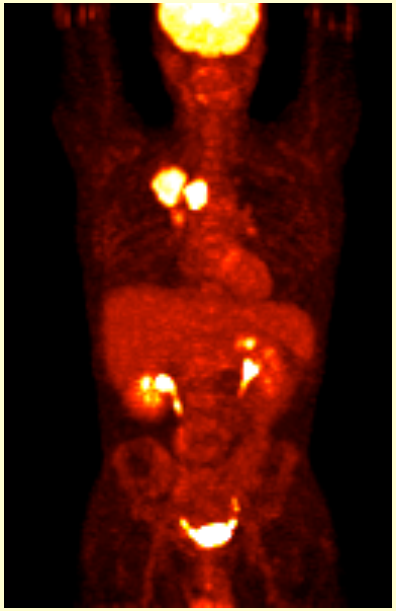
Exemple d'application 2 - Données

Images TEP au FDG

- Fantôme :
  - Acquis sur le TEP/TDM Gemini GXL (HEGP)
  - 6 réplicats d'une acquisition de 6 sphères de volume 0,5 à 16mL
  - 1 SUV par sphère, variant de 4,3 à 6,07 (mesuré à l'activimètre)



- Patients :
  - Acquis sur le TEP/TDM Discovery LS (Institut J. Bordet, Bruxelles)
  - 14 examens PET/CT
  - 32 tumeurs pulmonaires + métastases



**Tylski, Dusart, Garcia, Vanderlinden, Buvat SNM 2009 (soumis)*

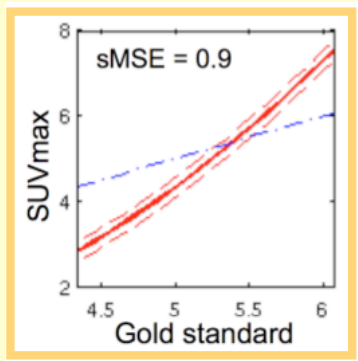
Exemple d'application 2 - Méthodes comparées

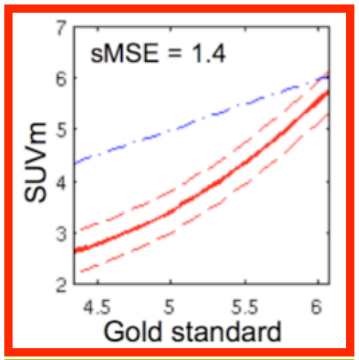
6 méthodes d'estimation de SUV :

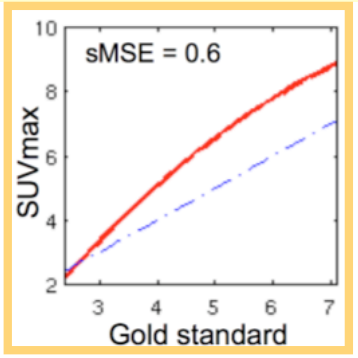
- SUV maximum : SUV_{max}
- SUV moyen dans une région fixe (cylindre de 15 mm de diamètre et de 15mm de hauteur) : SUV_f
- SUV moyen dans une région définie par un seuil d'intensité : SUV_m
- SUV moyen dans une région définie par un seuil d'intensité avec prise en compte de l'activité environnante : SUV_{mb}
- SUV moyen corrigé de l'effet de volume partiel par une méthode d'ajustement $SUV_{meanfit}$
- SUV moyen corrigé de l'effet de volume partiel par un coefficient de recouvrement SUV_{mRCt}

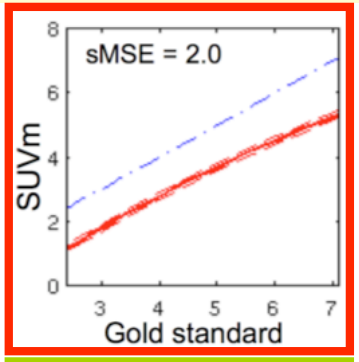
Mise en œuvre de GEWAGS pour comparer les 6 méthodes

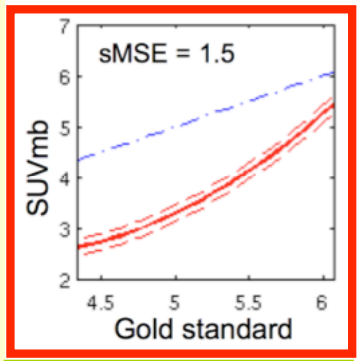
Exemple d'application 2 - Résultats

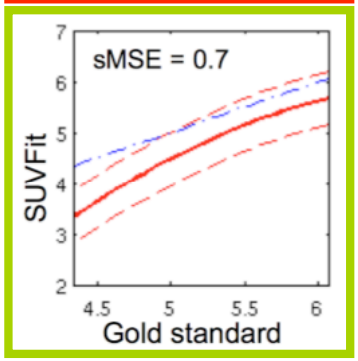


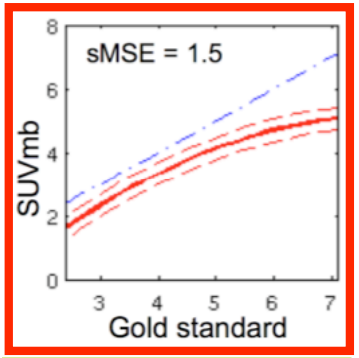


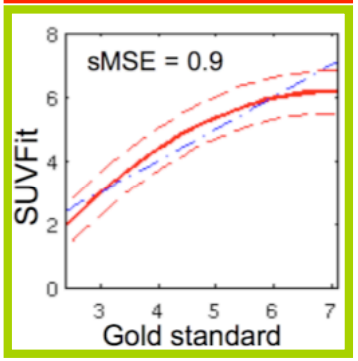


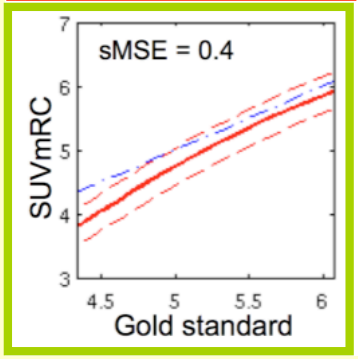


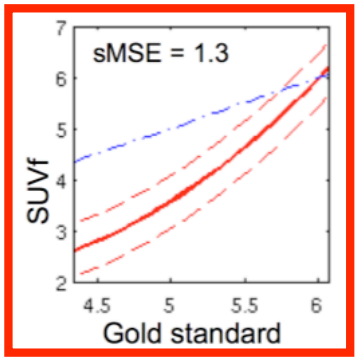


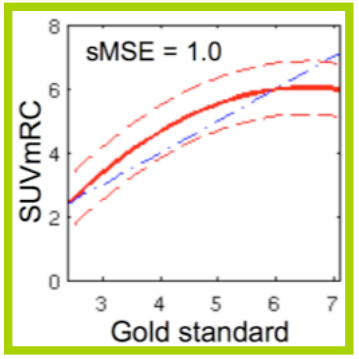


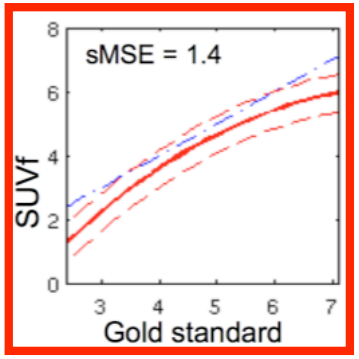












Données fantômes

Données cliniques

STAPLE

Simultaneous Truth and Performance Level Estimation

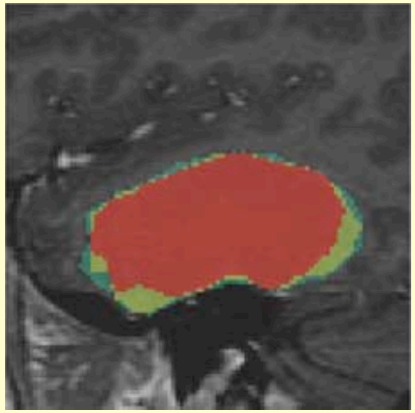
Idée un peu similaire, mais dédiée à la problématique de segmentation de structure(s)

Warfield, Zou, Wells, IEEE Trans Med Imaging 23: 903-921, 2004

STAPLE : hypothèses

Hypothèses :

1) Au moins 2 résultats de segmentation indépendants disponibles (m) : m décisions de segmentation dans chacun des N voxels : D (N,m)



IRM prostate

2) Segmentation vraie = variable binaire dans chaque voxel : T(N)

3) Chaque méthode de segmentation peut être complètement caractérisée par sa sensibilité p(m) et spécificité q(m)

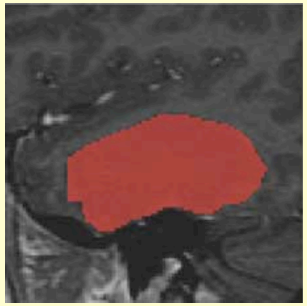
STAPLE : méthode

Méthodes :

1) Initialisation des p_m et q_m à la même valeur qqsoit m OU initialisation de la segmentation vraie

2) Maximisation itérative (EM) de la log-vraisemblance :

$$\mathcal{L}(\{D, T\} \mid \{p_m, q_m\}_{\text{pour tt m}})$$



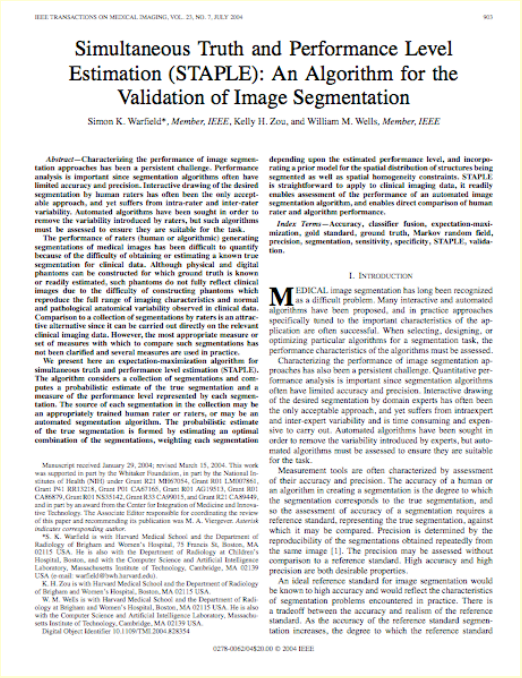
T

3) Caractérisation des performances de chaque méthode par p_m et q_m

Expert	$\hat{p}$	$\hat{q}$	PV
1	0.8951	0.9999	0.998
2	0.9993	0.9857	0.977
3	0.9986	0.9982	0.954

STAPLE : courte discussion

- Possibilité d'évaluer la segmentation de plusieurs structures
- Possibilité d'introduire des a priori (approche MAP)
- Possibilité d'introduire des contraintes spatiales
- Utilisé en IRM



IEEE TRANSACTIONS ON MEDICAL IMAGING, VOL. 23, NO. 7, JULY 2004 903

Simultaneous Truth and Performance Level Estimation (STAPLE): An Algorithm for the Validation of Image Segmentation

Simon K. Warfield*, *Member, IEEE*, Kelly H. Zou, and William M. Wells, *Member, IEEE*

***Abstract*—Characterizing the performance of image segmentation approaches has been a persistent challenge. Performance analysis is important since segmentation algorithms often have limited accuracy and precision. Interactive drawing of the desired segmentation by human raters has often been the only acceptable approach, and yet suffers from intra-rater and inter-rater variability. Automated algorithms have been sought in order to remove the variability introduced by raters, but such algorithms must be assessed to ensure they are suitable for the task.**

The performance of raters (human or algorithmic) generating segmentations of medical images has been difficult to quantify because of the difficulty of obtaining or estimating a known true segmentation for clinical data. Although physical and digital phantoms can be constructed for which ground truth is known or readily estimated, such phantoms do not fully reflect clinical images due to the difficulty of constructing phantoms which reproduce the full range of imaging characteristics and normal and pathological anatomical variability observed in clinical data. Comparison to a collection of segmentations by raters is an attractive alternative since it can be carried out directly on the relevant clinical imaging data. However, the most appropriate measure or set of measures with which to compare such segmentations has not been clarified and several measures are used in practice.

We present here an expectation-maximization algorithm for simultaneous truth and performance level estimation (STAPLE). The algorithm considers a collection of segmentations and computes a probabilistic estimate of the true segmentation and a measure of the performance level represented by each segmentation. The source of each segmentation in the collection may be an appropriately trained human rater or raters, or may be an automated segmentation algorithm. The probabilistic estimate of the true segmentation is formed by estimating an optimal combination of the segmentations, weighting each segmentation depending upon the estimated performance level, and incorporating a prior model for the spatial distribution of structures being segmented as well as spatial homogeneity constraints. STAPLE is straightforward to apply to clinical imaging data, it readily enables assessment of the performance of an automated image segmentation algorithm, and enables direct comparison of human rater and algorithm performance.

***Index Terms*—Accuracy, classifier fusion, expectation-maximization, gold standard, ground truth, Markov random field, precision, segmentation, sensitivity, specificity, STAPLE, validation.**

I. Introduction

MEDICAL image segmentation has long been recognized as a difficult problem. Many interactive and automated algorithms have been proposed, and in practice approaches specifically tuned to the important characteristics of the application are often successful. When selecting, designing, or optimizing particular algorithms for a segmentation task, the performance characteristics of the algorithms must be assessed.

Characterizing the performance of image segmentation approaches has also been a persistent challenge. Quantitative performance analysis is important since segmentation algorithms often have limited accuracy and precision. Interactive drawing of the desired segmentation by domain experts has often been the only acceptable approach, and yet suffers from intraexpert and inter-expert variability and is time consuming and expensive to carry out. Automated algorithms have been sought in order to remove the variability introduced by experts, but automated algorithms must be assessed to ensure they are suitable for the task.

Measurement tools are often characterized by assessment of their accuracy and precision. The accuracy of a human or an algorithm in creating a segmentation is the degree to which the segmentation corresponds to the true segmentation, and so the assessment of accuracy of a segmentation requires a reference standard, representing the true segmentation, against which it may be compared. Precision is determined by the reproducibility of the segmentations obtained repeatedly from the same image [1]. The precision may be assessed without comparison to a reference standard. High accuracy and high precision are both desirable properties.

An ideal reference standard for image segmentation would be known to high accuracy and would reflect the characteristics of segmentation problems encountered in practice. There is a tradeoff between the accuracy and realism of the reference standard. As the accuracy of the reference standard segmentation increases, the degree to which the reference standard

Manuscript received January 29, 2004; revised March 15, 2004. This work was supported in part by the Whitaker Foundation, in part by the National Institutes of Health (NIH) under Grant R21 MH67054, Grant R01 LM007861, Grant P41 RR13218, Grant P01 CA67165, Grant R01 AG19513, Grant R01 CA86879, Grant R01 NS35142, Grant R33 CA99015, and Grant R21 CA89449, and in part by an award from the Center for Integration of Medicine and Innovative Technology. The Associate Editor responsible for coordinating the review of this paper and recommending its publication was M. A. Viergever. *Asterisk indicates corresponding author.*

*S. K. Warfield is with Harvard Medical School and the Department of Radiology of Brigham and Women's Hospital, 75 Francis St, Boston, MA 02115 USA. He is also with the Department of Radiology at Children's Hospital, Boston, and with the Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA 02139 USA (e-mail: warfield@hwh.harvard.edu).

K. H. Zou is with Harvard Medical School and the Department of Radiology of Brigham and Women's Hospital, Boston, MA 02115 USA.

W. M. Wells is with Harvard Medical School and the Department of Radiology at Brigham and Women's Hospital, Boston, MA 02115 USA. He is also with the Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA 02139 USA.

Digital Object Identifier 10.1109/TMI.2004.828354

Warfield, Zou, Wells, IEEE Trans Med Imaging 23: 903-921, 2004

Conclusion

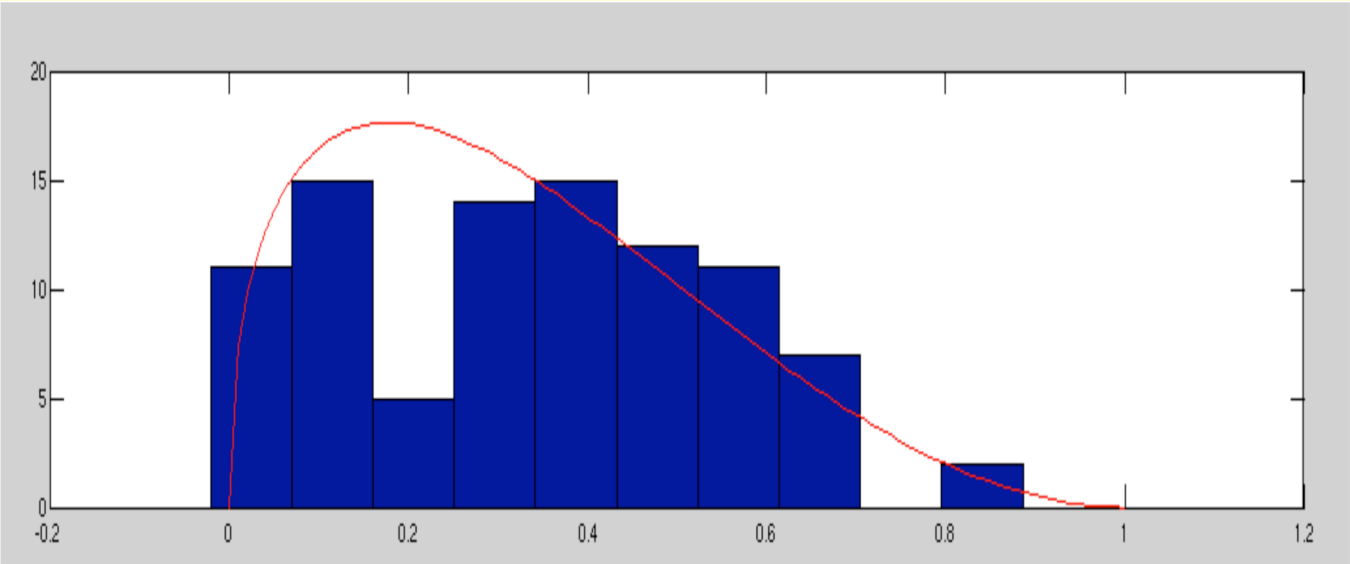
• Une évaluation rigoureuse est possible même en l'absence de gold standard

Diaporama disponible sur
http://www.guillemet.org/irene/conferences

Exemple de distribution beta

Patients simulés de l'étude MA / DCL

nombre de valeurs



PL normalisé

Cohérence des données réelles et simulées (étude MA - DCL)