

Comment évaluer les méthodes de détermination de volumes métaboliquement actifs en Tomographie par Emission de Positons (TEP) ?

Irène Buvat

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

buvat@imnc.in2p3.fr
<http://www.guillemet.org/irene>

Problématique

Plus de 20 méthodes ont été proposées pour délimiter le volume tumoral métaboliquement actif :

- seuil fonction du pourcentage du max (quel pourcentage ?)
- seuil fonction du pourcentage du max et de l'activité environnante (quelle fonction ?) Nestle et al 2005, David et al 2006, Nestle et al 2007
- seuil ajusté itérativement (quelle fonction ?) Daisne et al 2003, Jentzen et al 2007, Drever et al 2007, van Dalen et al 2007, Black et al 2004, Green et al 2008
- méthodes de détection de contours Geets et al 2007, Drever et al 2007, Li et al 2008
- méthodes de classification de voxels Hatt et al 2007, Zhu et al 2003, Aristophanous et al 2007, Hatt et al 2009, Montgomery et al 2007

Quelles méthodes sont les plus fiables dans telle et telle circonstances ?

Première option : méta-analyse de la littérature (1)

	Etude	Seuil	Données optimisation	Données validation	Performances
Variété des données	Erdi et al, 1997	- fonction du volume anatomique	Fantôme cylindrique Sphères : 0,4 à 5,5 mL RCA : 2,8 3,1 5,5 7,4	17 tumeurs	$E_{VTDM} = 8,44 \pm 7,66\%$
	Nestle et al, 2005	- fonction de l'activité de fond - fonction de l'activité max - fixe (SUV = 2,5)	Non précisé	25 tumeurs	Seuil fonction de l'activité de fond plus « stable » que les autres méthodes
	Schaefer et al, 2008	- fonction de l'activité de fond	Fantôme cylindrique Sphères : 7 à 258 mL RCA : 2,5-33	Fantôme cylindrique Sphères : 7 à 171mL RCA : 3 5 10 15 tumeurs	sphères : $-0,4 < E_R < 0,7\text{mm}$ tumeur : $-0,7 < E_{RTDM} < 1,2\text{mm}$
Variété de l'optimisation	Davis et al, 2006	-fonction de l'activité de fond	Fantôme cylindrique Sphères : 0,04 à 47 mL RCA : 3 10 20	30 tumeurs	$E_{RTDM} < 1,4 \text{ mm}$
	Daisne et al, 2003	- fonction du RCA	Fantôme cylindrique Sphères : 0,5 à 17,5mL RCA : 1,5- 8,7		Faible influence de la reconstruction sur le seuil optimal
	Jentzen et al, 2007	- fonction du RCA	Fantôme cylindrique Sphères : 0,5 à 26,5 mL RCA : 2 3 5 9	Fantôme cylindrique Sphères : 0,5 à 26mL RCA : 10 16 tumeurs $^{18}\text{F-FDG}$ 23 tumeur ^{124}I	$E_{VTDM} = 10\%$ pour des sphères $> 10\text{mL}$ $E_{VTDM} = 9\%$ (resp 15%) pour des tumeurs $< 7,5 \text{ mL}$ (resp $>$)
Variété des figures de mérite	Drever et al, 2007 b	- fonction du RCA coupe par coupe (2D) - fonction du RCA 3D	Fantôme cylindrique Sphères : 6,4 22,4 107,5mL RCA : 2 - 15	Idem étalonnage + 2 inserts non sphériques dans un fantôme cylindrique	sphères : $E_V -36 \text{ à } 80\%$ (2D), -6 à 22% (3D) inserts : -61 à 7%(2D), -6 à 18% (3D)
	Green et al, 2008	- fonction du SUVmoyen (croissance de région)	Images analytiques RCA : 2 4 8	31 tumeurs entre deux examens	Validation dans le cadre du suivi avec des paramètres calculés dans la région segmentée
	Black et al, 2004	- fonction du SUVmoyen, déterminé par régression (Treg) - seuil : 42% du SUVmax (T42%)	Fantôme cylindrique Sphères : 2 à 291 mL RCA : 4 - 21	15 tumeurs	T42% : $E = -23\%$ (sphères) Erreur relative entre T42% et Treg entre -93 et -35%

Première option : méta-analyse de la littérature (2)

Variété des données

Variété de l'optimisation

Variété des figures de mérite

	Etude	Méthodes	Données optimisation	Données validation	Performances
Variété des données	Geets et al, 2007	- Gradient (G) - Seuil de Daisne et al (T_D)	Inconnu	Simulations analytiques : fantôme cylindrique, sphères 4,2 à 51 mL, RCA : 6	-3,17 < E_V < 1,81%
				fantôme cylindrique physique, sphères 2,1 à 92,9 mL, RCA : 1,5-15	-20 < E_V < -10%
				7 tumeurs (ORL) avec échantillon chirurgical	G : -10 < $E_{V\text{chir}}$ < 46% T_D : 8 < $E_{V\text{chir}}$ < 191%
Variété de l'optimisation	Drever et al, 2007 a	- Seuil fonction du fond - Filtre Sobel - Gradient	Fantôme cylindrique : Cylindres de Ø 12, 25 et 47 mm et sphères de 6,4 22,4 107,5mL, RCA : 2-15		Seule la méthode de seuil permet de retrouver correctement le volume des structures
Variété des figures de mérite	Hatt et al 2007	-Classification floue avec Champ de Markov (FHMC) - seuil 42% max (T42%)	Inconnu	Fantôme cylindrique simulé (MC) , Sphères 0,5 à 26mL, RCA 4 et 8: FS	pour $V > 1,2 \text{ mL}$: FHMC : $E_{\text{class}} < 25\%$ T42% : $10 < E_{\text{class}} < 200\%$
			Inconnu	Fantôme physique FP même caractéristique que FS	pour $V > 1,2 \text{ mL}$: FHMC: $E_{\text{class}} < 30\%$ T42% : $20 < E_{\text{class}} < 120\%$
Variété des figures de mérite	Hatt et al, 2009	-Classification floue sans Champ de Markov(FLAB) -c moyennes floues (FCM) -FHMC -T42%	Fantôme Simulé (FS), identique à celui utilisé dans Hatt et al, 2007 pour optimiser FLAB	- FP, identique à FP dans Hatt et al, 2007	pour $V > 1,2 \text{ mL}$: FLAB : $5 < E_{\text{class}} < 15\%$ pour $V > 2,6 \text{ mL}$ FHMC : $5 < E_{\text{class}} < 30\%$ FCM : $5 < E_{\text{class}} < 38\%$ pour $V > 5,6 \text{ mL}$ T42% : $8 < E_{\text{class}} < 60\%$
				- Tumeurs hétérogènes simulées	FLAB : $-6 < E_{\text{class}} < 10\%$ FHMC : $-8 < E_{\text{class}} < 15\%$ FCM : $-30 < E_{\text{class}} < 30\%$

Première option : méta-analyse de la littérature (3)

Variété des données
Variété de l'optimisation
Variété des figures de mérite

Etude	Méthodes	Données optimisation	Données validation	Performances
Li et al, 2008	- Contours actifs - Seuil de Jentzen et al	Sphères de F1 avec RCA : 10	F1 : Fantôme cylindrique Sphères : 0,5 à 20mL, RCA : 2 à 16	$-79 < E_V < -6,2 \%$
Zhu et al, 2003	-C moyennes floues (CMF)	Inconnu	Tumeurs du cerveau (Nombre inconnu)	Faisabilité de la méthode
Dewalle-Vignon et al, 2008	-CMF sur projections (Vretro) - Seuil T40% du max	Images simulées	10 images de patients	Indice de Jaccard entre Vretro et T40% de 0,80 (+/-0,08)
Aristophanous et al, 2007	- Modèle de mélange gaussien (GMM) - Seuil 40% du max - Seuil de Nestle et al	3 tumeurs	4 tumeurs	GMM faisable GMM donne toujours des volumes supérieurs aux deux autres méthodes
Montgomery et al, 2007	- C moyennes (CM) - Champ de Markov (MRFM) -MRFM multi échelle (MRFM-ME)	Inconnu	Fantôme anthropomorphe 4 ellipsoïdes en cire de 1,4 à 3,1 mL 10 patients	CM : $-12,9 < E_V < -6\%$ MRFM $-4,4 < E_V < 9,5\%$ MRFM-ME $-7,31 < E_V < -2$ Validation visuelle sur les images de patients

Première option : méta-analyse de la littérature (4)

Discussion : impossibilité de comparer les performances des différentes approches uniquement sur la base des données de la littérature du fait de :

- la variété des données utilisées
- la variété des approches d'optimisation des paramètres
- la variété des figures de mérite

Nécessité de comparer les méthodes à partir des mêmes données, en utilisant la même approche d'optimisation et les mêmes figures de mérite pour statuer sur leurs performances respectives

Quelles données ?

Données pour l'évaluation

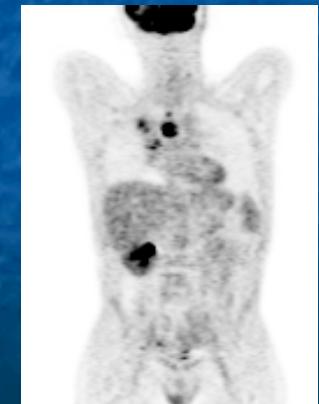
- Fantômes :
 - + accessibles à tous
 - trop simples (sphères homogènes, parois de plexiglas)



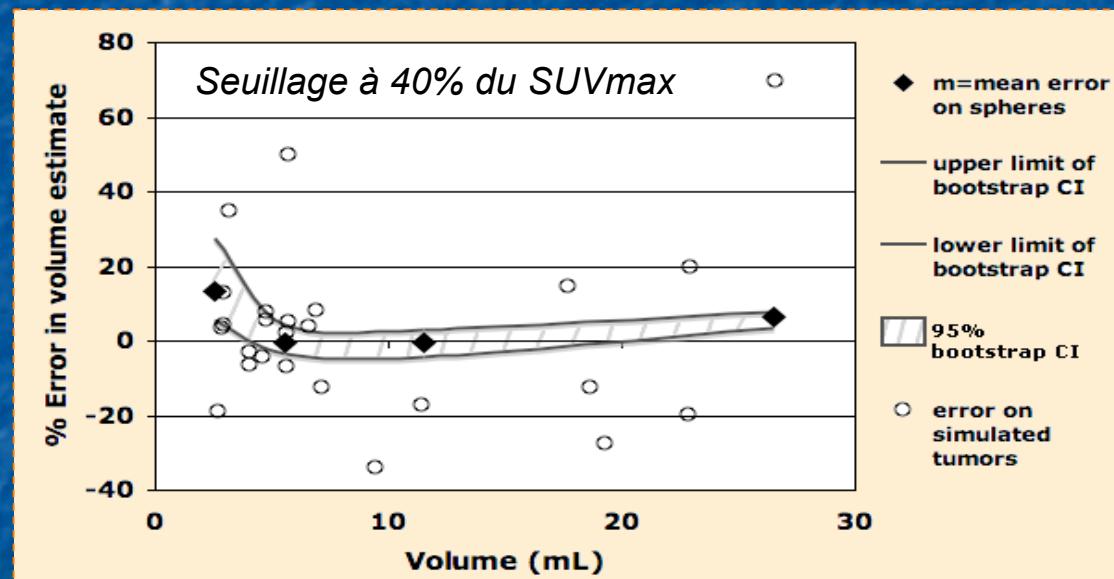
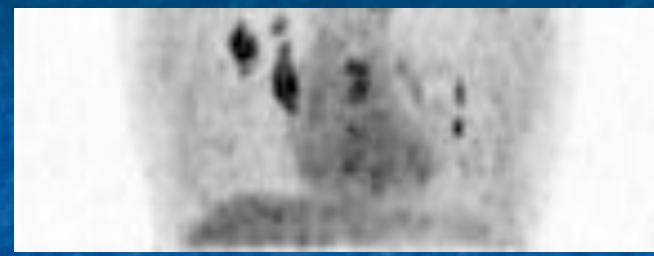
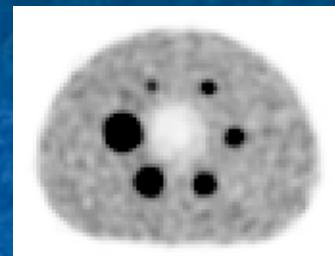
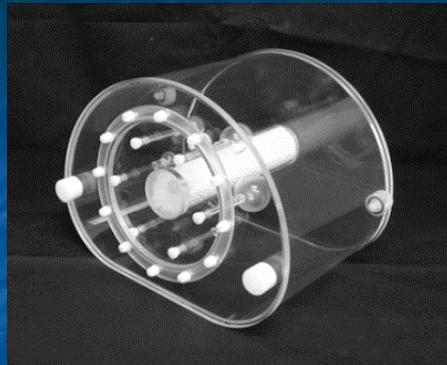
- Données simulées
 - pas forcément simples à générer
 - + possiblement très réalistes (tumeurs non sphériques, hétérogènes)



- Données cliniques
 - + représentatives de la réalité clinique
 - quel gold standard ?



Impact des données sur lesquelles on évalue



Les performances des méthodes dépendent **fortement** des données sur lesquelles elles opèrent

Optimisation des méthodes à évaluer (1)

La plupart des méthodes implique un ou plusieurs paramètres :

Approches	Méthodes	Travaux	Nb de paramètres
Intensité	Seuil	Erdi et al., 1997 Nestle et al., 2005 Mah et al., 2002 Davis et al., 2006 Schaefer et al., 2008 ...	1 ou 2 qui définissent le seuil
	Seuil itératif	Daisne et al., 2003 Jentzen et al., 2007; Drever et al., 2007b Black et al., 2004	Au moins 2 paramètres (si une seule courbe)
	Seuil avec croissance de région	Green et al., 2008	Seuil
Forme	Gradient	Geets et al., 2007 Drever et al., 2007a	Paramètres des filtres de débruitage et déconvolution et de l'algorithme de classification des régions*
	Contour fermé	Li et al., 2008	Paramètres de l'énergie interne (contour) et externe (image)
Classification	C-moyennes floues	Zhu et Jiang, 2003 Hatt et al, 2009 Dewalle et al, 2008	Nombre de classes* Initialisation
	Mélange gaussien sur l'histogramme	Aristophanous et al., 2007	Nombre de classes* Initialisation
	Modèle statistique et champs de Markov	Montgomery et al., 2007 Hatt et al., 2007	Nombre de classes*, paramètres du modèle statistique, et des champs de Markov*
	Modèle statistique	Hatt et al., 2009	Nombre de classes*, paramètres du modèle statistique

Optimisation des méthodes à évaluer (2)

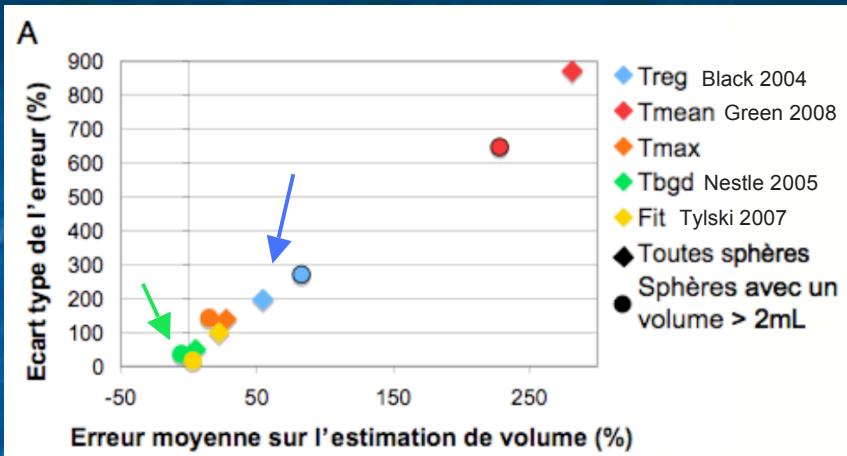
Comment fixer ces paramètres ?

Plusieurs stratégies :

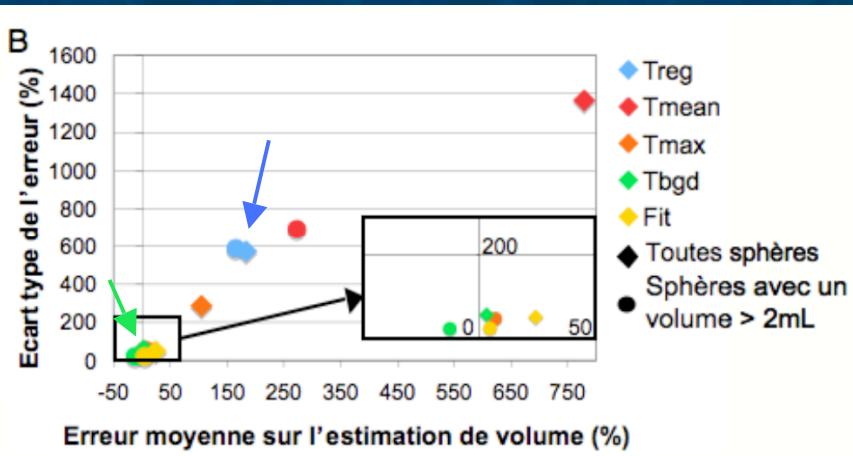
- prendre les valeurs de la littérature
 - non optimisées pour le type de données à traiter,
pénalise la méthode
- calibrer avec un fantôme
 - + le plus réaliste
- optimiser sur les données à traiter
 - biaisé

Les performances des méthodes dépendent de la façon
dont les paramètres sont optimisés

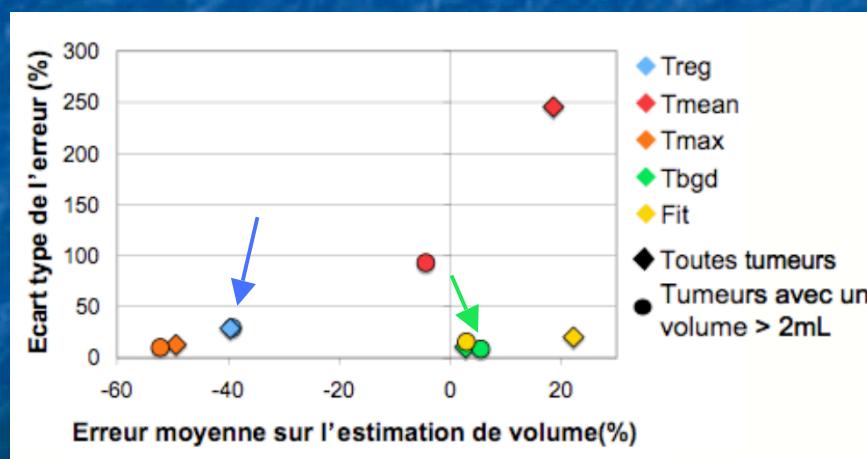
Impact de la méthode d'optimisation des méthodes à évaluer



Optimisation leave-one out :
Méthodes donnant leurs meilleurs résultats



Optimisation sur un échantillon d'apprentissage et évaluation sur un échantillon test



Optimisation sur un fantôme (tumeurs sphériques) et évaluation sur un patient simulé

Figure de mérite à considérer

- Le plus souvent : erreur sur le volume

$$\frac{100 * (\text{volume estimé} - \text{volume vrai})}{\text{volume vrai}}$$

Adapté pour la caractérisation des tumeurs ou de leur évolution

- Pour la définition de volume cible, une erreur de classification est plus pertinente

Faux positif (FP) = volume identifié comme appartenant à la tumeur à tort

Faux négatif (FN) = volume tumoral non identifié comme tel

Daisne et al, Radiology 2004

Erreur de classification = $(FP + FN)/\text{volume tumoral}$

Hatt et al, Phys Med Biol 2007

Pistes à considérer pour aller plus loin dans l'évaluation

- Caractériser les performances conjointement sur différents types de données
- Bases de données partagées, fantômes, simulées ou cliniques
- Evaluation clinique sans gold standard

Caractériser les performances sur différents types de données

Les résultats sont différents suivant le type de données

Erreurs sur le volume

Simulations analytiques :

-3,2% (4,2 ml) à +1,8% (51 ml)

Fantômes :

-20% (2,1 ml) à -10% (92,9 ml)

Patients (vs volume de la pièce chirurgicale) :

+15% (4,1 ml) à -10% (30,9 ml)
mais +47% (17,3 ml)*

Geets et al, EJNMMI 2007

Fantômes :

11,6% (7,4 ml) à -3,9% (171,3 ml)

Patients (vs volume TDM) :

+1,3% (66,4 ml) à -27,8% (12,7 ml)
mais 0% (5,5 ml)*

Schaefer et al, EJNMMI 2008

Fantômes :

> 50% (0,48 ml) à <10% (26,52 ml)

Patients (vs volume TDM) :

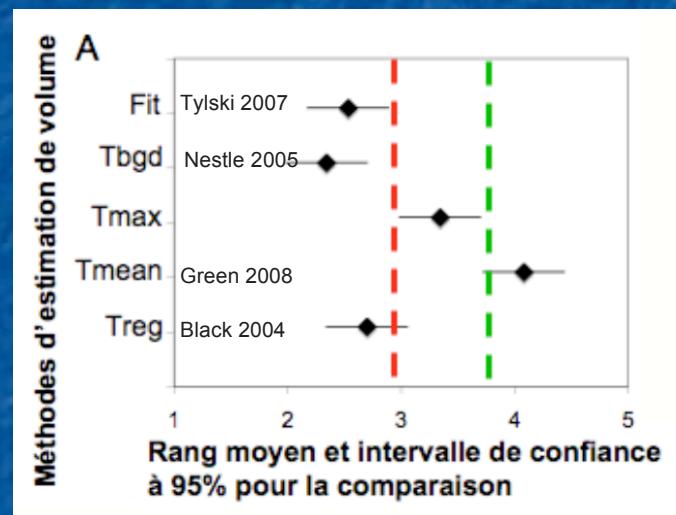
~20% (~2,5 ml) à -25% (17.5 ml)*

Jentzen et al, JNM 2007

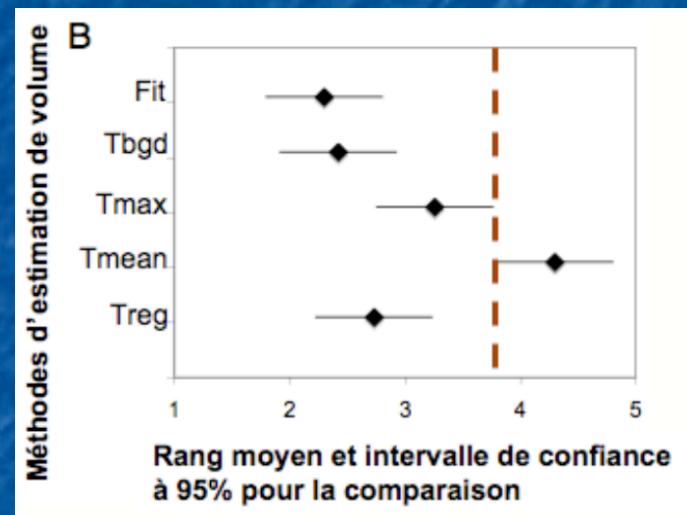
* Pas d'évolution monotone de l'erreur en fonction du volume de la tumeur

Caractériser les performances sur différents types de données

Option : commencer par classer les méthodes plutôt que de chercher à caractériser leurs performances absolues, pour dégager une cohérence



Optimisation leave-one out :
Méthodes donnant leurs meilleurs résultats

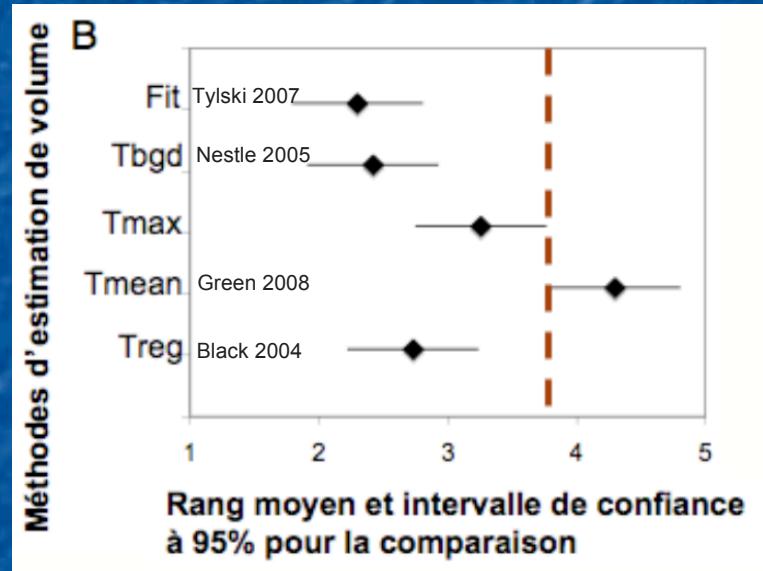


Optimisation sur un échantillon
d'apprentissage et évaluation sur un
échantillon test

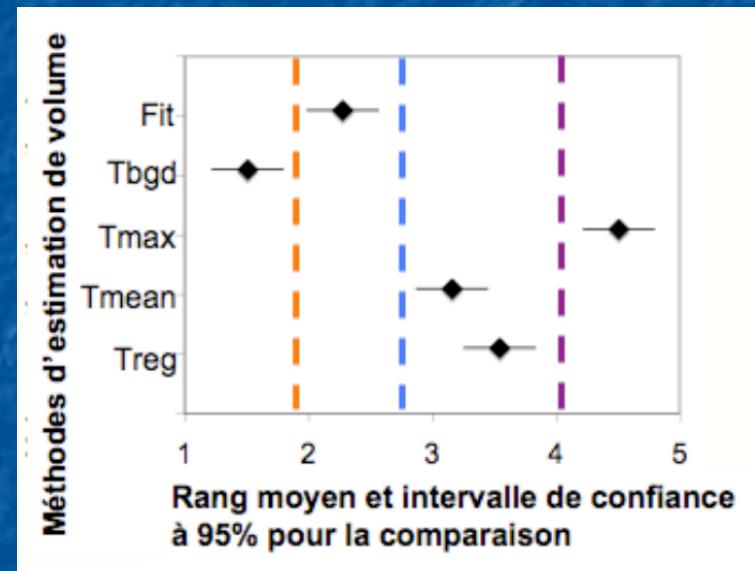
Moindre dépendance à la façon dont les méthodes sont optimisées si
l'optimisation des différentes méthodes est consistante (toutes
optimisées suivant le même critère)

Caractériser les performances sur différents types de données

Révèle aussi la robustesse des méthodes vis-à-vis de différents jeux de données



Optimisation sur un échantillon d'apprentissage et évaluation sur un échantillon test, fantôme



Optimisation sur un fantôme et évaluation sur des données cliniques simulées

Caractériser les performances sur différents types de données

Exemple : comparaison de 5 méthodes de délimitation de volumes sur les mêmes données (issues de fantômes variés) et avec différentes procédures d'optimisation

Comparative Assessment of Methods for Estimating Tumor Volume and Standardized Uptake Value in ^{18}F -FDG PET

Perrine Tylski, Simon Stute, Nicolas Grotus, Kaya Doyeux, Sébastien Hapdey, Isabelle Gardin, Bruno Vanderlinden, and Irène Buvat

¹IMNC UMR 8165 CNRS–Paris 7 and Paris 11 Universities, Orsay, France; ²LITIS EA 4108 Laboratory, University of Rouen, Rouen, France; and ³Nuclear Medicine Department, Bordet Institute, Université Libre de Bruxelles, Brussels, Belgium

... mais beaucoup d'autres méthodes prometteuses mériteraient d'être testées, et des données encore plus réalistes (avec mvt respiratoire et fixations hétérogènes) mériteraient d'être considérées

Base de données partagées

Motivation : étendre la comparaison de méthodes incluant celles qu'on n'a pas reprogrammées

- Construire des bases de données suffisamment représentatives : fantômes, données simulées, données cliniques



Nécessité de fournir les données afférentes permettant d'optimiser les méthodes

Action dans le cadre du GDR (recenser puis compiler) ?

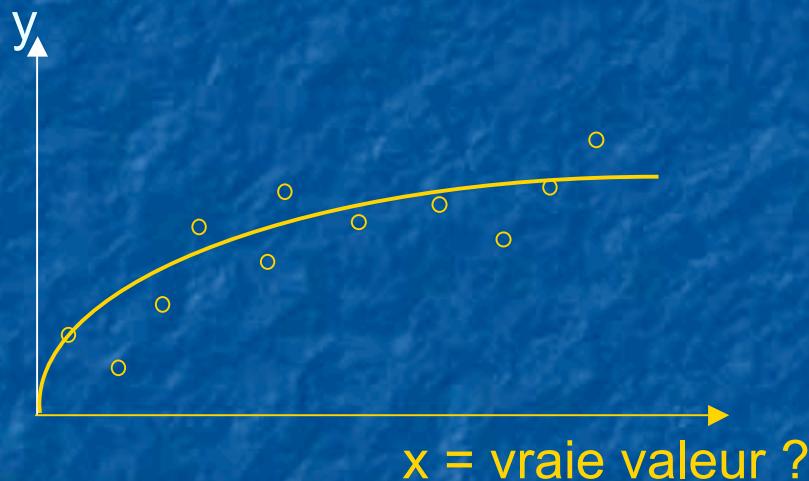
Evaluation clinique

- Par rapport à des spécimens chirurgicaux : caractériser la fiabilité de la référence d'abord
- Evaluations sans gold standard qui méritent d'être explorées davantage
 - GEWAGS
 - STAPLE

Evaluation sans gold standard (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 $m = 1, 2, \dots$ 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

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

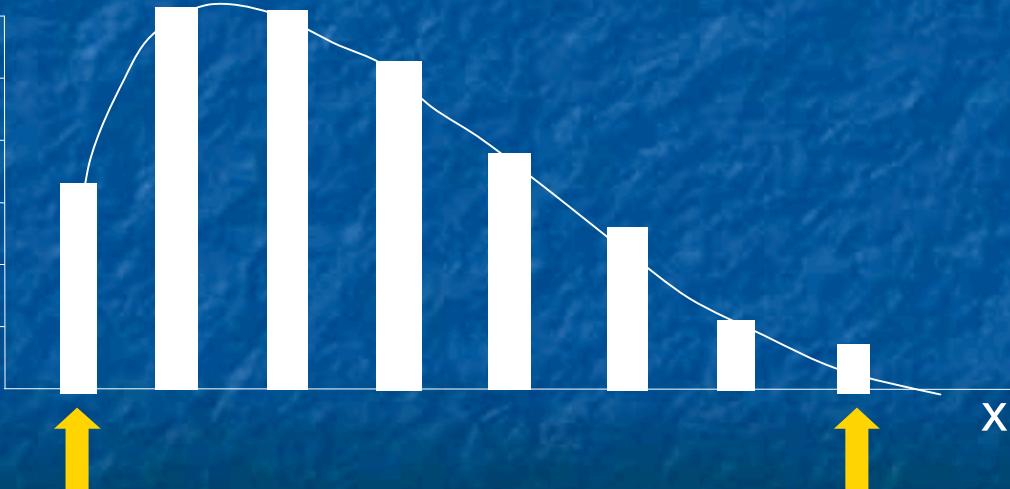
Evaluation sans gold standard (2)

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

nombre de valeurs



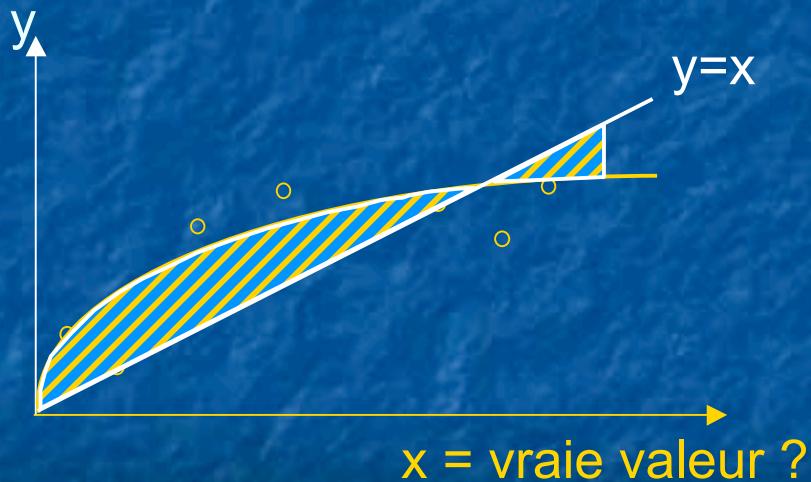
Evaluation sans gold standard (3)

Méthode :

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

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

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



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

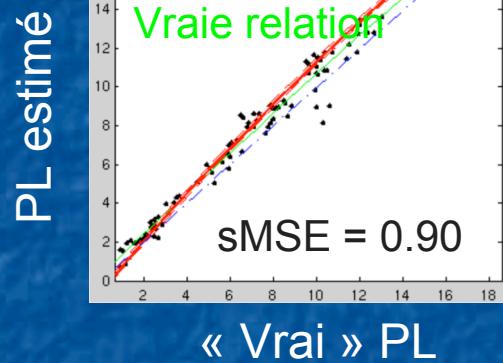
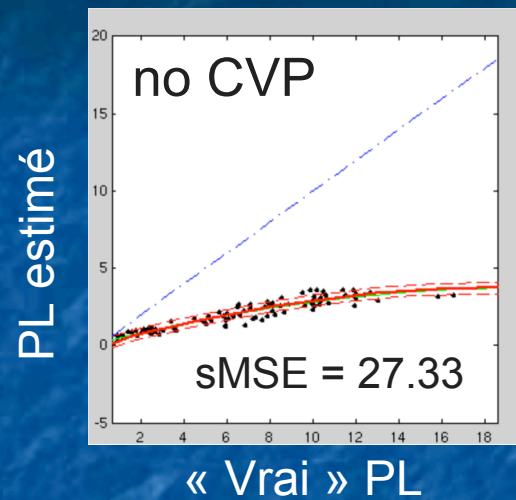
Evaluation sans gold standard (4)

- 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*

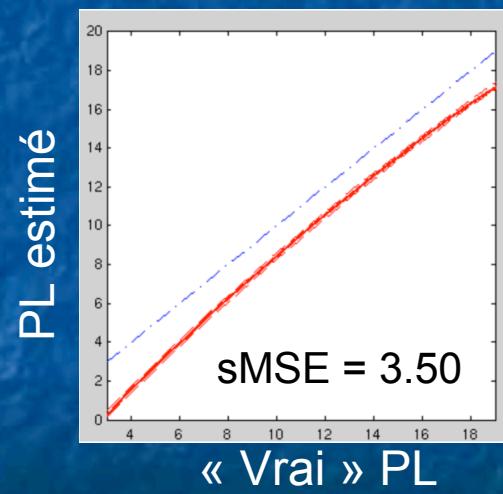
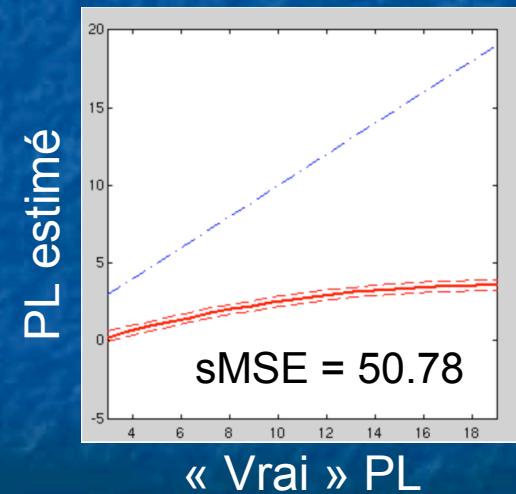
Exemple d'application : mesure du potentiel de liaison en SPECT

Patients simulés



p<0.01

Patients réels



p<0.01

Soret et al. J Nucl Med 44: 1184-1193, 2003

Soret et al. Nucl Instrum Meth Phys Res A, 2007

Evaluation sans gold standard

- STAPLE

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, weighing 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 intra-expert 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 MH067054, Grant R01 LM007961, Grant R01 RR13218, Grant P01 CA67663, Grant R01 RR19513, Grant R01 CA86879, Grant R01 NS35142, Grant R01 CA99443, and Grant R21 CA99449, and in part by the National Institute for Integration of Macromolecular Informative 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 at 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@wh.harvard.edu).

K. H. Zou is with Harvard Medical School and the Department of Radiology at 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

0278-0062/04\$20.00 © 2004 IEEE

Conclusions

- Nécessité de mener des études comparatives des méthodes de segmentation des volumes cibles pour clarifier leurs performances respectives
- L'évaluation doit inclure la caractérisation des biais ET la robustesse (variabilité des résultats) en fonction de l'optimisation et des données traitées
- Tendre vers l'évaluation sur des données réalistes voire réelles

Remerciements

Perrine Tylski

Institut Jules Bordet, Bruxelles
Université Catholique de Louvain, Bruxelles

