

Action Spécifique ICoMIM

## Evaluation en imagerie médicale : méthodologie et outils

Irène Buvat  
U494 INSERM, Paris

buvat@imed.jussieu.fr

30 janvier 2003

Action Spécifique ICoMIM - Irène Buvat - janvier 2003 - 1

L'évaluation est un travail complexe et souvent difficile à mener de façon très rigoureuse. Mon objectif aujourd'hui est de vous donner d'une part, un cadre méthodologique qui permet de bien situer les différents travaux d'évaluation auxquels on peut être confronté, d'autre part, de vous indiquer les outils qui existent pour mener à bien un travail d'évaluation de façon objective et rigoureuse.

## Evaluation : objectifs



Action Spécifique ICoMIM - Irène Buvat - janvier 2003 - 2

Tout d'abord, voyons pourquoi on cherche à évaluer.

Objectif de l'évaluation
Déterminer <u>le système d'imagerie ou la technique d'analyse d'images</u> qui conduit à <u>la meilleure image ou à l'information la plus pertinente</u> <u>dans un contexte donné</u>
Quoi évaluer ?
Dans quel but ?
Sur quelles données ?
<small>Action Spécifique ICoMIM - Irène Buvat - janvier 2003 - 3</small>

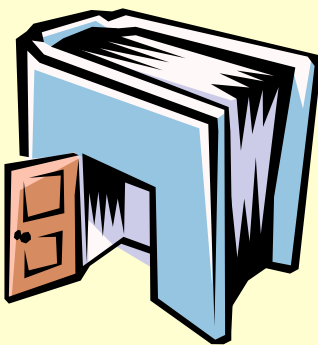
Typiquement, on fait appel à des méthodes d'évaluation pour répondre à des questions du type : «Quel est le système d'imagerie ou la technique d'analyse d'images qui conduit à la meilleure image ou à l'information la plus pertinente dans un contexte donné ?».

Cette seule formulation révèle déjà l'ambiguïté du processus d'évaluation. En effet, elle met en évidence la nécessité de préciser d'emblée :

- ce qu'on cherche à évaluer
- la question à laquelle on cherche à répondre
- et le contexte dans lequel on souhaite répondre à cette question.

Pour clarifier un peu les différentes situations que l'on rencontre dans les travaux d'évaluation, une approche conceptuelle rigoureuse aux problèmes d'évaluation a été développée depuis plus d'une dizaine d'années.

## Terminologie



Cette approche repose sur une terminologie bien définie, que je vais maintenant vous présenter.

Qu'évalue t-on ?



Système

*E.g., dispositif d'imagerie (nouvel imageur)  
méthode de reconstruction tomographique  
méthode de segmentation*

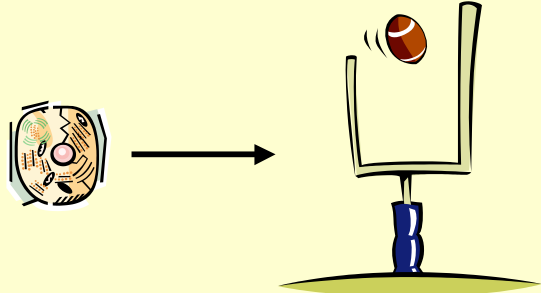
Action Spécifique ICoMIM - Irène Buvat - janvier 2003 - 5

Tout d'abord, il faut bien sur commencer par définir ce qu'on veut évaluer. Typiquement, il s'agit soit d'un dispositif d'imagerie ou de méthodes de synthèse ou d'analyse d'images.

On peut désigner ce qu'on cherche à évaluer sous le nom générique du système.

Par exemple, le système pourra être un nouveau type de détecteur, une méthode de reconstruction tomographique, ou une méthode de segmentation d'images.

Pour quelle finalité ?



Tâche

- Détection ou classification binaire, e.g., *absence ou présence de lésion lésion bénigne ou maligne*
- Estimation, e.g., *valeur de fraction d'éjection, taux de métabolisme de glucose*

Action Spécifique ICoMIM - Irène Buvat - janvier 2003 - 6

La deuxième notion qu'il est indispensable de préciser dans un travail d'évaluation est l'information que l'on cherche à extraire des images. De façon générique, c'est ce qu'on appelle la tâche.

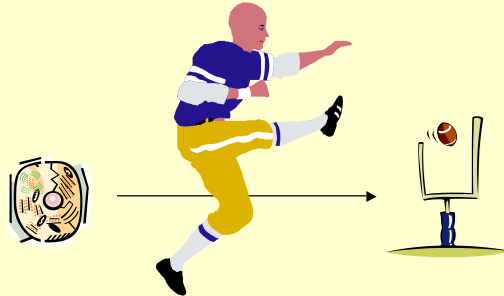
De façon générale, on peut distinguer 2 types de tâches : les tâches de détection et les tâches d'estimation.

Les tâches de détection sont les tâches qui appellent généralement une réponse binaire : présence ou absence d'anomalies.

Les tâches d'estimation sont celles qui conduisent à l'estimation d'un paramètre sur une échelle continue, comme une fraction d'éjection, un flux sanguin.

Il est fondamental de bien distinguer ces deux types de tâches, et de savoir définir le type de tâche au centre du problème d'évaluation, car les méthodes d'évaluation à mettre en œuvre vont ensuite totalement dépendre de la tâche.

Qui effectue la tâche ?



Observateur

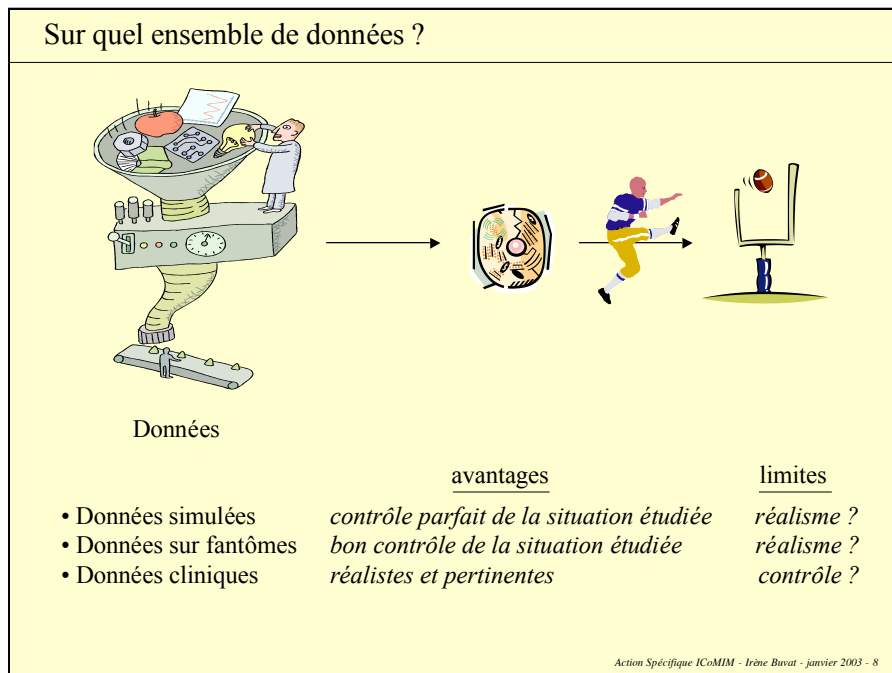
- Observateur humain *e.g., radiologue*
- Observateur numérique *i.e., algorithme*

Action Spécifique ICoMIM - Irène Buvat - janvier 2003 - 7

La troisième notion importante est qui va effectuer la tâche : l'entité qui va effectuer la tâche est appelée, de façon générique, l'observateur.

L'observateur peut être de 2 types :

- il peut s'agir d'un observateur humain, comme un radiologue qui va interpréter les images,
- ou d'un observateur numérique, c'est à dire un algorithme qui va estimer une valeur quantitative de fraction d'éjection par exemple.



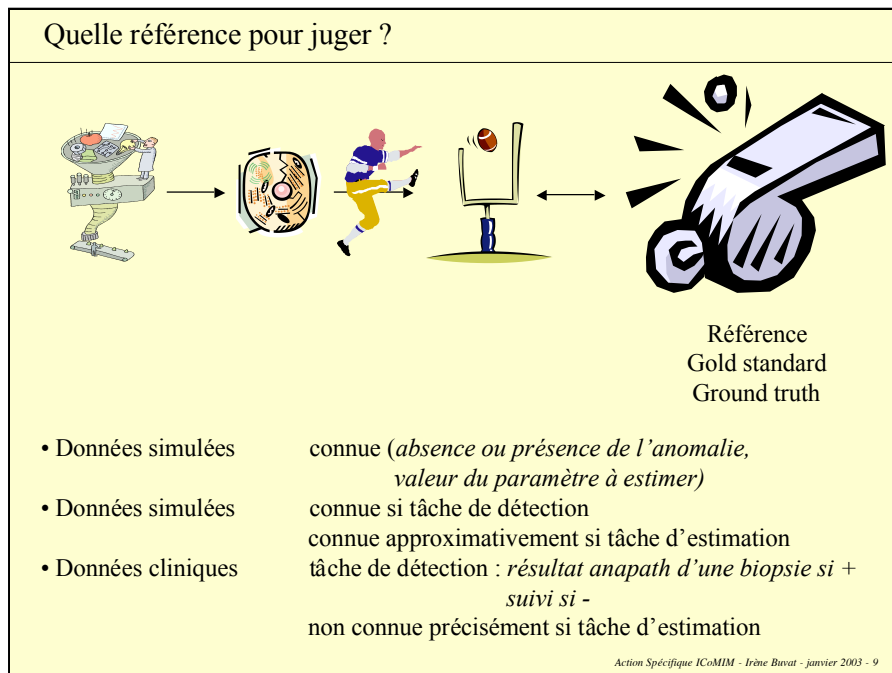
Outre la tâche et les observateurs, il faut aussi déterminer l'ensemble de données sur lequel on va réaliser le travail d'évaluation.

Il peut s'agir de données simulées numériquement : pour ces données, l'investigateur a un contrôle total des situations auxquelles il s'intéresse. En revanche, le réalisme des données simulées peut être mis en question.

Il peut s'agir de données acquises sur des fantômes physiques. Là encore, l'avantage de ces données est qu'elles sont généralement bien contrôlées, aux erreurs expérimentales près. Cependant, comme pour les données simulées, la limite concomitante est le réalisme des configurations qui peuvent être étudiées au moyen d'un fantôme.

In fine, le travail d'évaluation devrait se faire dans des conditions proches des conditions dans lesquelles le système évalué est censé opérer, c'est à dire sur des données cliniques. Nous allons voir pourquoi on ne privilégie pas systématiquement ce type de données, qui sont pourtant a priori les plus pertinentes.





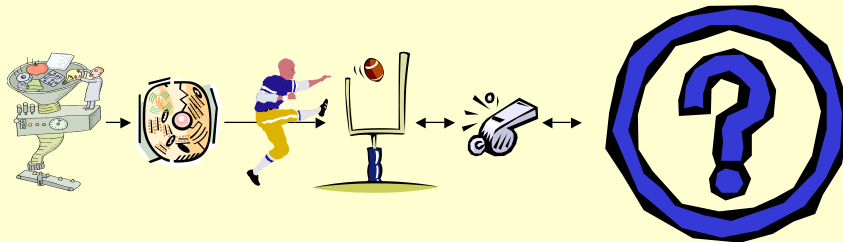
Une autre notion fondamentale en évaluation est celle de la référence, encore appelée gold standard ou ground truth. Ce qu'on appelle ainsi, c'est la vraie réponse, c'est à dire la vraie valeur du paramètre que l'on cherche à évaluer par exemple, ou l'absence ou la présence d'une lésion dans les tâches de détection.

Quand on travaille sur des données simulées, le gold standard est parfaitement connu. C'est par exemple la valeur du paramètre que l'on a simulé, ou le fait que l'on a simulé une anomalie ou pas.

Dans le cas de données sur fantômes, la situation est un peu plus complexe que sur des données simulées, du fait des erreurs expérimentales inévitables. Mais dans l'ensemble, dans ces configurations, le gold standard reste bien connu.

Enfin, les difficultés de définition d'un gold standard apparaissent essentiellement quand on travaille sur des données cliniques. Pour les tâches de quantification, en général, le paramètre que l'on cherche à évaluer n'est pas précisément connu. Pour les tâches de détection, là encore, un contrôle par biopsie est envisageable en présence d'une lésion, mais la définition du gold standard est beaucoup plus difficile à établir lorsque les images ne révèlent pas de lésion.

A quelle question répondre ?



Question posée

- Figure de mérite            résultat chiffré  
   e.g., *erreur moyenne de  $X \pm Y\%$*
  - Test d'hypothèse            appelant une réponse binaire  
   e.g., *le filtrage A conduit à des meilleures performances de détection d'anomalies que le filtrage B*
- passage obligé par une figure de mérite

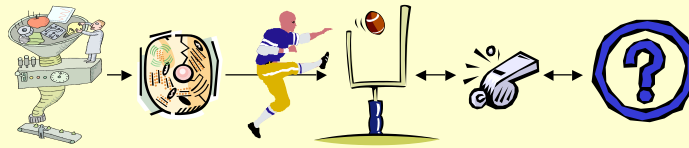
Action Spécifique ICoMM - Irène Buvat - janvier 2003 - 10

Enfin, l'objectif de tout travail d'évaluation est d'apporter une réponse à une question. Cette question se formule soit comme l'estimation d'une figure de mérite, lorsque la question appelle une réponse chiffrée, soit comme un test d'hypothèse, lorsqu'elle appelle à une réponse binaire (hypothèse acceptée ou rejetée).

En fait, les tests d'hypothèses sous-tendent systématiquement une figure de mérite.

C'est donc davantage l'usage que l'on fait de la figure de mérite qui distingue ces deux cas. Soit on l'interprète telle qu'elle, soit on l'utilise pour tester une hypothèse appelant une réponse binaire.

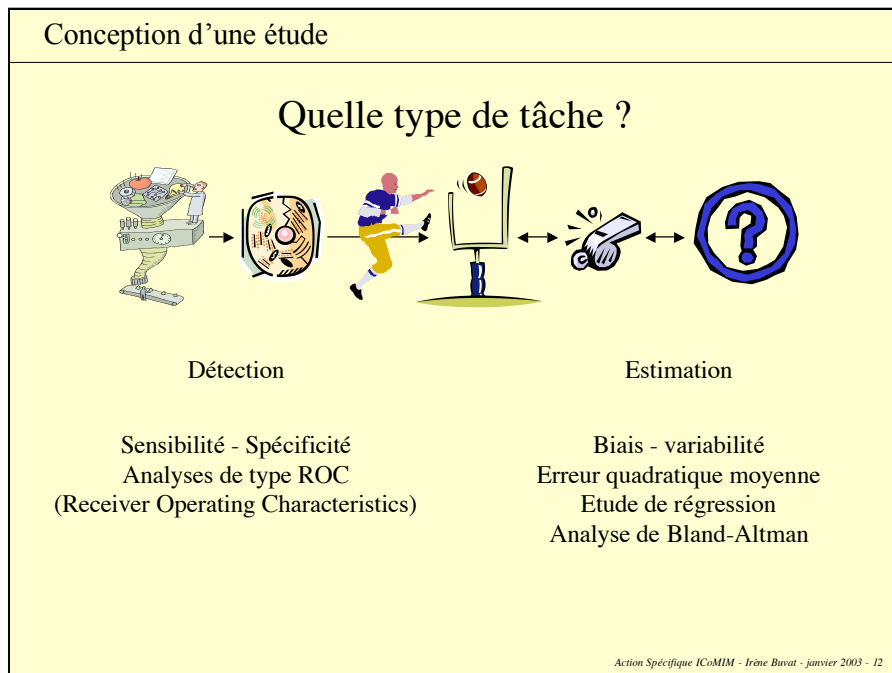
## Conception d'une étude



Action Spécifique ICoMIM - Irène Buvat - janvier 2003 - 11

Ces notions étant définies, il faut identifier les composantes qui vont orienter le choix de la méthodologie d'évaluation adaptée pour répondre à la question que l'on se pose.

C'est ce que nous allons voir maintenant.



La composante qui permet d'orienter le choix de la méthodologie à adopter est la tâche au centre du travail d'évaluation : tâche de détection ou tâche d'estimation.

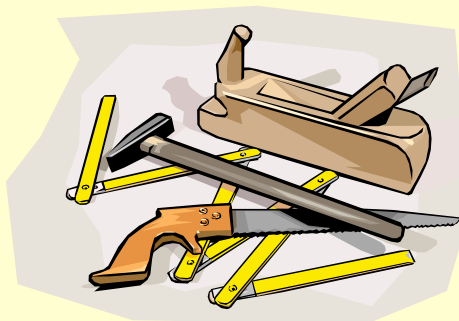
En simplifiant, pour toutes les tâches de détection, les méthodes d'évaluation adaptées sont celles qui vont consister à faire des calculs de sensibilité et de spécificité. La méthodologie la plus générale et la plus performante pour traiter la majeure partie des problèmes d'évaluation impliquant une tâche de détection est celle reposant sur le concept d'analyse ROC, pour Receiver Operating Characteristics.

A l'opposé, pour les tâches d'estimation, les approches appropriées sont celles qui font appel à des calculs de biais, de variabilité des mesures, voire d'erreur quadratique moyenne. Dans certains cas, il est pertinent de faire des études de régression ou des études de type Bland-Altman.

Donc en résumé, le type de tâche détermine totalement le type de méthodologie d'évaluation qu'il va falloir mettre en œuvre.

Nous allons maintenant voir plus en détails ces approches, et dans quels contextes elles sont pertinentes.

## Outils pour l'évaluation



*Action Spécifique ICoMIM - Irène Buvat - janvier 2003 - 13*

Voyons donc les outils disponibles pour l'évaluation.

## Outils pour les tâches de détection

Sensibilité - spécificité - exactitude



Approche ROC



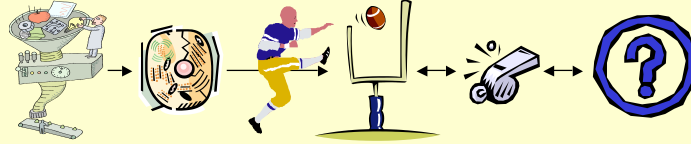
Action Spécifique ICoMIM - Irène Buvat - janvier 2003 - 14

Nous allons commencer par parler des outils à mettre en œuvre pour les tâches de détection.

Dans ce contexte, une approche souvent utilisée consiste à faire des calculs de sensibilité et de spécificité de détection. Je vais montrer pourquoi cette approche est restrictive.

Nous verrons ensuite comment on peut aller au delà de cette approche, en utilisant la méthodologie ROC.

## Tâches de détection : contexte général



Données divisibles en 2 catégories

*e.g., avec et sans anomalie :  
positif (avec) ou négatif (sans)*

Tâche : déterminer la catégorie à laquelle appartient chaque élément du jeu de données



Observateur :  
- humain (e.g., radiologue interprétant l'image)  
- algorithmique (algorithme calculant un indice et classant l'image à partir de la valeur de l'indice)

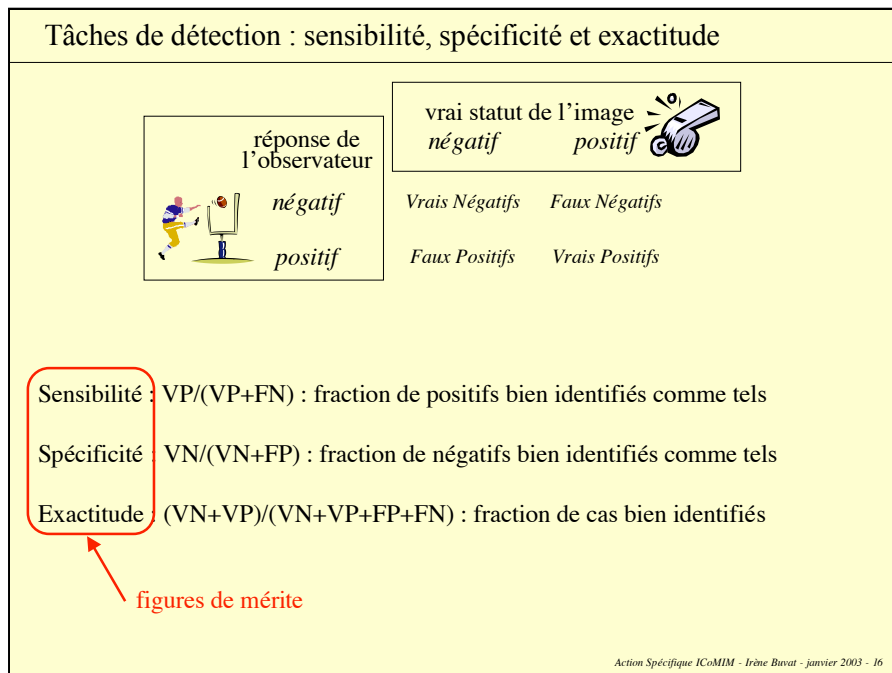
Action Spécifique ICoMM - Irène Buvat - janvier 2003 - 15

Tout d'abord, pour comprendre les outils, précisons le contexte des tâches de détection.

Les données peuvent être divisées en 2 catégories : par exemple, certains éléments comportent une anomalie, d'autres pas.

La tâche assignée à l'observateur est donc de déterminer, pour chaque élément, à laquelle des 2 catégories il appartient, c'est à dire s'il contient ou non l'anomalie. C'est donc une décision binaire.

Pour ces tâches, l'observateur peut être humain ou algorithmique. Un exemple d'observateur algorithmique serait un algorithme qui calculerait un indice à partir d'une image, par exemple une fraction d'éjection, et qui, en fonction de la valeur de l'indice, classerait l'image comme contenant ou non une anomalie.



L'analyse des réponses des observateurs va se faire en triant les réponses négatives et positives en fonction de la vraie réponse. Ceci conduit à séparer 4 catégories d'images :

- les images positives et identifiées comme telles : on parle de vrais positifs.
- les images négatives et identifiées comme telles : ce sont les vrais négatifs.
- les images positives classées comme négatives : ce sont les faux négatifs.
- enfin, les images négatives classées comme positives : ce sont les faux positifs.

Les effectifs dans ces 4 catégories permettent de calculer des index caractérisant la justesse de la classification, et en particulier :

- la sensibilité, définie comme le pourcentage de cas positifs correctement identifiés.
- et la spécificité, définie comme le pourcentage de cas négatifs correctement identifiés.

On utilise parfois aussi l'exactitude, c'est à dire le pourcentage de cas, positifs ou négatifs, correctement identifiés.

A priori, ces grandeurs, qui sont des figures de mérite, pourraient être suffisantes pour évaluer une méthode de détection. En fait, elles présentent très vite des limites, dont nous allons voir des exemples maintenant.



## Pourquoi cette approche est insuffisante ? Deux exemples



Comment évaluer le compromis sensibilité/spécificité ?

e.g. : méthode A : sensibilité = 80% spécificité = 65%

méthode B : sensibilité = 90% spécificité = 50%

Quelle est la meilleure méthode ?

Insuffisance de l'exactitude pour répondre :

méthode A : sensibilité = 70%, spécificité = 90%, exactitude = 87%

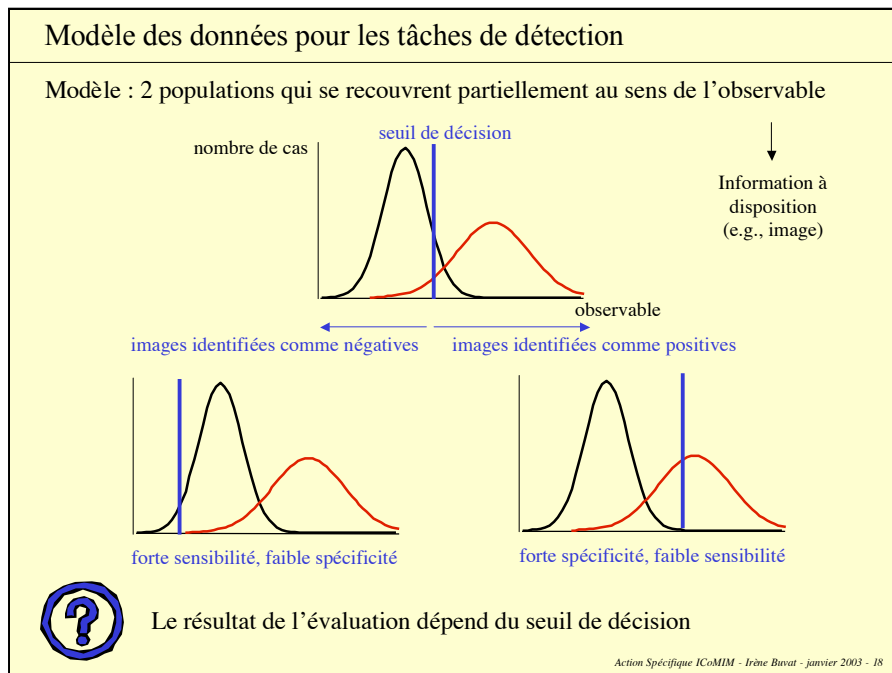
méthode B : sensibilité = 40%, spécificité = 96%, exactitude = 87%

Une même exactitude peut correspondre à des compromis sensibilité/spécificité très différents

Action Spécifique ICoMM - Irène Buvat - janvier 2003 - 17

La première limite est que pour répondre à la question que l'on se pose concernant les performances de la méthode de détection, il faut généralement prendre en compte à la fois sa sensibilité et sa spécificité. Or, on arrive souvent à des situations du type de celle-ci. Quelle est alors dans ce cas la meilleure des 2 méthodes ?

On pourrait penser que l'exactitude apporte une réponse. Cependant, cet exemple montre que ce n'est pas le cas. En effet, une même exactitude peut correspondre à des compromis sensibilité/spécificité bien différents, et la considération de l'exactitude sans considérer les couples sensibilité et spécificité peut conduire à des conclusions erronées.



Pour mieux analyser et dépasser les limites de la sensibilité et de la spécificité, il faut considérer un modèle des données. Les données peuvent être représentées comme étant issues de 2 populations, correspondant aux 2 catégories de données, qui se recouvrent partiellement au sens de l'observable. L'observable correspond à l'information à disposition, c'est-à-dire typiquement une image.

L'observateur, pour donner une réponse binaire à la question qui lui est posée, se fixe inconsciemment ce qu'on appelle un seuil de décision. Au delà de ce seuil de décision, il qualifie les images de positives, et en dessous, il est considéré comme négatives.

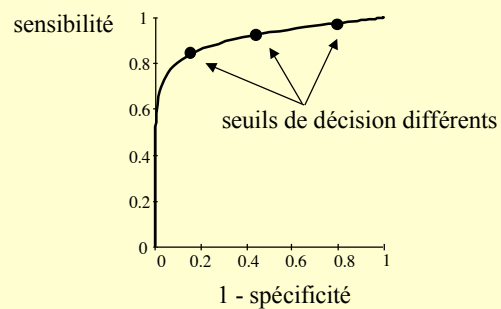
Plus ce seuil est bas, plus la sensibilité va être élevée, mais plus la spécificité sera faible. Et inversement.

Ce modèle met clairement en évidence le fait que les résultats, en termes de sensibilité et de spécificité, dépendent totalement du seuil de décision. En cela, l'évaluation n'est donc pas satisfaisante puisque les résultats sont susceptibles de dépendre fortement de l'observateur et du seuil de décision qu'il utilise.

## Solution : l'approche ROC



Caractérisation des performances de détection indépendante d'un seuil de décision

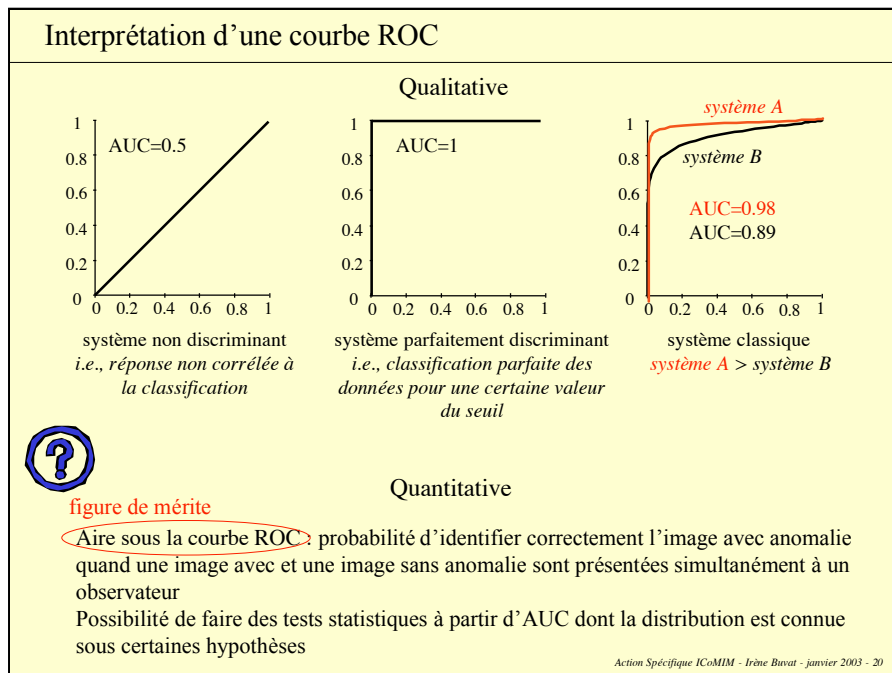


Action Spécifique ICoMM - Irène Buvat - janvier 2003 - 19

Les approches ROC apportent une solution à ce problème de seuil de décision puisque tout leur intérêt est de permettre l'évaluation d'un système indépendamment du choix d'un seuil de décision.

Plus précisément, l'approche ROC consiste à représenter la valeur de la sensibilité en fonction de (1-spécificité) pour toutes les valeurs de seuil possibles, et à joindre ces points par une courbe. Chaque point de la courbe représentant le compromis sensibilité/spécificité correspondant à un seuil de décision spécifique.

La courbe ROC résume l'ensemble des compromis sensibilité/spécificité pour les différentes valeurs de seuil de décision.



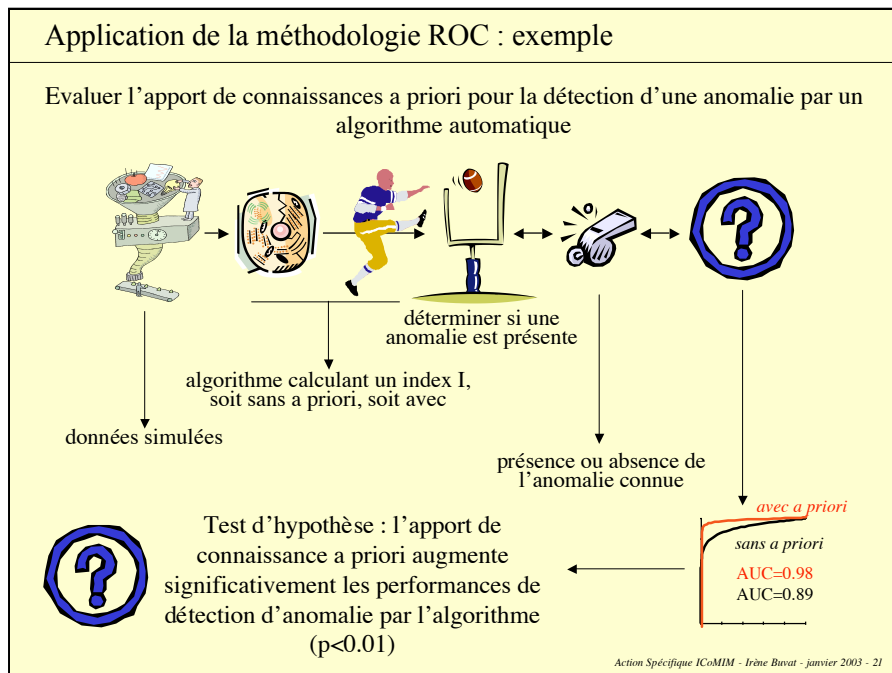
L'interprétation d'une courbe ROC peut se faire de manière qualitative. La forme de la courbe caractérise en effet simplement les performances de détection.

Une courbe confondue avec la diagonale correspond à un système non discriminant, c'est à dire que la réponse donnée par l'observateur n'est nullement corrélée à la présence ou à l'absence d'une anomalie.

Une courbe de cette forme correspond à un système parfaitement discriminant, pour lequel il existe un seuil de décision séparant parfaitement les deux sous-groupes.

Dans la pratique, on rencontre des courbes de forme intermédiaire. Par exemple ici, la courbe correspondant au système A révèle à des performances de détection meilleure que celle correspondant au système B.

Lorsqu'on veut réaliser un test d'hypothèse, ou classer les méthodes par performances croissantes, il est nécessaire d'extraire une figure de mérite. La figure de mérite associée aux courbes ROC est l'aire sous la courbe ROC, qui représente la probabilité d'identifier correctement l'image avec anomalie lorsqu'une image avec anomalie et une image sans sont présentées simultanément à l'observateur.



Cette méthodologie ROC s'applique à de nombreux problèmes d'évaluation. On peut l'illustrer sur un exemple qui serait celui de l'évaluation de l'apport de connaissances a priori pour la détection d'un type d'anomalie.

Dans ce cas, la tâche consisterait à déterminer si l'anomalie est absente ou présente.

Le système à évaluer serait en fait confondu avec l'observateur : il s'agirait de l'algorithme calculant un index I. Deux versions de l'algorithme seraient testées : l'une sans et l'autre avec apport de connaissances a priori.

Imaginons que l'on applique ces deux algorithmes à des données simulées, pour lesquelles la présence ou l'absence de l'anomalie est donc parfaitement connue.

Pour chacun des algorithmes, on va calculer une courbe ROC.

La superposition des deux courbes ROC donne déjà des indications sur la potentielle supériorité d'une méthode par rapport à une autre.

Les aires sous les courbes ROC peuvent être calculées comme figure de mérite et permettre de conclure, par un test statistique approprié, à la supériorité effective de l'une ou l'autre méthode.

## Potentialités de l'approche ROC



Détection et localisation des anomalies : extensions LROC, FROC et AFROC

Anomalies multiples par images : extensions FROC et **AFROC** *Utilise un logiciel ROC classique*

Observateurs multiples : approche Dorfman-Berbaum-Metz (1992)

Absence de gold standard : travaux de Henkelman (1990) et Beiden (2000)

Exploitation de l'appariement si les mêmes images sont traitées par deux méthodes à comparer (augmentation de la puissance statistique des tests)

Données continues (index issu d'un algorithme) ou discrètes (scores attribués par des radiologues)

Modèle paramétrique (2 lois normales) ou non paramétrique de l'observable

Compatible avec des observateurs humains ou algorithmiques

Action Spécifique ICoMIM - Irène Buvat - janvier 2003 - 22

L'approche ROC a été adaptée à de nombreuses configurations pratiques, que j'ai tenté de répertorier sur cette diapositive.

Des variantes de l'approche ROC traitent des cas où l'on s'intéresse non seulement à la détection, mais aussi à la localisation de l'anomalie. Ce sont les approches LROC pour Localized ROC, les approches FROC pour Free ROC et AFROC pour Alternative Free ROC.

Ces deux dernières approches permettent de traiter des cas où plusieurs anomalies sont potentiellement présentes dans l'image. L'approche AFROC en particulier permet d'utiliser un logiciel d'analyse ROC classique pour effectuer l'analyse de données FROC.

L'approche ROC permet de gérer des résultats obtenus au moyen d'observateurs multiples.

Quelque chose qui est insuffisamment connu est que une méthode a également été proposée pour mettre en œuvre une analyse ROC en l'absence de gold standard.

Pour augmenter la puissance des analyses ROC, on peut tirer parti de l'appariement de données, par exemple lorsque chaque image est traitée de deux façons différentes avant d'être interprétée et que l'on veut comparer les deux traitements.

L'approche ROC est adaptée que les données soient discrètes (scores attribués par des observateurs à des images) ou continues (index issu d'un algorithme).

Je vous ai indiqué que le modèle sous-jacent à l'approche ROC était un modèle de loi binormale. En fait, l'approche ROC peut aussi s'appliquer quand ce modèle n'est pas vérifié. On parle d'études ROC non paramétriques dans ce cas.

Enfin, l'approche ROC n'implique pas nécessairement des observateurs humains. Comme dans l'exemple que j'ai montré tout à l'heure, elle peut parfaitement s'appliquer à l'évaluation d'algorithmes.

## En pratique...



Nombreux programmes disponibles en ligne :

<http://www.bio.ri.ccf.org/Research/ROC/>

<http://www-radiology.uchicago.edu/krl/toppage11.htm#software>

<http://www.radiology.arizona.edu/krupinski/mips/rocprog.html>

Bibliographie sur :

<http://www.guillemet.org/lirene/equipe4/ressources.html>

Action Spécifique ICoMIM - Irène Buvat - janvier 2003 - 23

En pratique, il existe de nombreuses ressources permettant de mettre en œuvre l'approche ROC sans avoir à reprogrammer quoi que ce soit. Je vous invite en particulier à visiter ces sites Web où vous trouverez les liens vers les pages où des programmes peuvent être téléchargés.

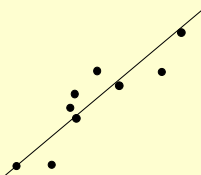
Enfin, des références bibliographiques figurent sur cette page Web, pour vous permettre d'en savoir plus sur toutes les approches ROC.

## Outils pour les tâches d'estimation

Biais et variabilité



Corrélation



Action Spécifique ICoMIM - Irène Buvat - janvier 2003 - 24

Passons maintenant aux outils nécessaires à l'évaluation de tâches d'estimation.

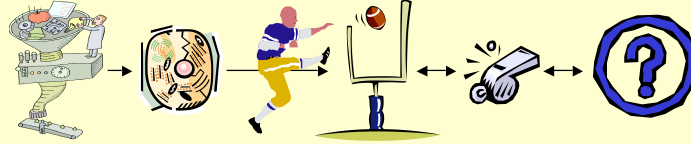
A priori, l'évaluation est plus facile. En pratique pourtant, les méthodes simples qui existent peuvent conduire à des interprétations erronées.

Je vais donc revenir plus en détails sur les deux type d'outils dont on dispose pour évaluer des tâches d'estimation :

- d'abord, les notions de biais et de variabilité.
- ensuite, les notions de corrélation.



## Tâches d'estimation : contexte général



Données quelconques

Tâche : extraire une valeur à partir d'une image



Observateur :  
- algorithmique  
- humain et algorithmique (valeur calculée à partir d'une ROI tracée manuellement)

Action Spécifique ICoMIM - Irène Buvat - janvier 2003 - 25

Tout d'abord, resituons le contexte. Il s'agit cette fois d'extraire une valeur à partir d'une image ou d'une série d'images.

Les données peuvent bien sur être des images simulées, acquises sur fantômes ou des données cliniques.

L'estimation peut être réalisée directement par un algorithme, ou par la un observateur aidé d'un algorithme : par exemple, ce peut-être la valeur moyenne dans une ROI tracée manuellement par un observateur.

Le travail d'évaluation va consister à déterminer la fiabilité des valeurs extraites.

## Tâches d'estimation : biais



Requièrent la connaissance de la vraie valeur du paramètre  
Applicables à des données simulées ou acquises sur fantômes  
seulement



Figure de mérite : biais  $\pm$  écart-type  
Tests d'hypothèse possibles

Plusieurs mesures du biais possibles :

$$\% \text{ erreur} = 1/N[\sum_{\text{observations}_i} (p_{i\_estimé} - p_i) / p_i]$$

$$\% \text{ erreur absolue} = 1/N[\sum_{\text{observations}_i} |p_{i\_estimé} - p_i| / p_i]$$

$$\text{erreur quadratique moyenne} = 1/N[\sum_{\text{observations}_i} (p_{i\_estimé} - p_i)^2 / p_i^2]$$

à choisir en fonction du contexte

Action Spécifique ICoMM - Irène Buvat - janvier 2003 - 26

L'approche la plus intuitive va consister à calculer l'erreur affectant le paramètre estimée, c'est-à-dire le biais.

Une limite à cette approche est qu'elle nécessite bien sur de connaître la vraie valeur du paramètre, c'est-à-dire qu'elle ne s'applique en toute rigueur qu'à des données synthétiques ou des données acquises sur fantômes.

Si on est capable d'estimer un biais, on peut bien sur l'utiliser directement comme figure de mérite, et on peut également faire des tests d'hypothèses, par exemple pour déterminer si les biais obtenus pour deux méthodes concurrentes sont similaires ou si une méthode conduit à des biais significativement plus faibles que l'autre.

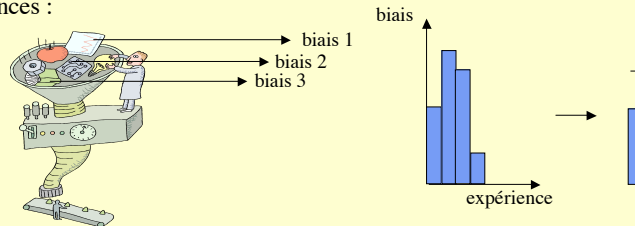
En fonction du contexte, plusieurs index sont possibles pour caractériser le biais, comme le pourcentage d'erreur, le pourcentage d'erreur absolue, ou l'erreur quadratique moyenne. La mesure la plus pertinente dépend du contexte.

## Tâches d'estimation : variabilité



Tout résultat en terme de biais doit être accompagné d'une estimation de la variabilité du biais pour être interprétable

Attention, la variabilité doit être calculée à partir d'un grand nombre d'expériences :



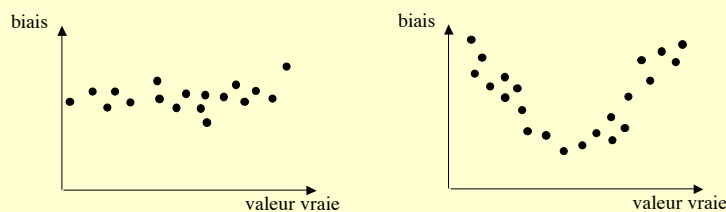
et pas au sein d'une même ROI (hypothèse d'ergodicité généralement non vérifiée, Buvat et al, J Nucl Med 42:101P, 2001 )

Action Spécifique ICoMM - Irène Buvat - janvier 2003 - 27

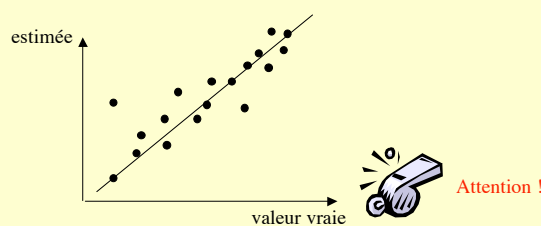
Attention, il est indispensable d'assortir le biais d'une mesure de la variabilité du biais. Cette mesure de variabilité ne peut se faire qu'en répétant un grand nombre de fois l'expérience, pour en déduire le biais moyen et l'écart-type du biais par exemple.

Il est impossible de déduire la variabilité du biais à partir d'une seule mesure. Une démarche parfois utilisée consiste à estimer la variabilité du biais à partir de la variabilité du signal dans la ROI servant à estimer le paramètre. Cette approche n'est pas légitime, et peut conduire à une estimation totalement erronée du biais, car l'hypothèse d'ergodicité, selon laquelle la moyenne spatiale est identique à la moyenne temporelle, n'est généralement jamais vérifiée sur des images médicales.

## Insuffisance des mesures de biais et variabilité



Evaluation plus complète : la régression linéaire ?



Action Spécifique ICoMIM - Irène Buvat - janvier 2003 - 28

Le biais et la variabilité sont des index très synthétiques, mais ils peuvent masquer une dépendance du biais à la vraie valeur du paramètre qu'il peut être important de connaître.

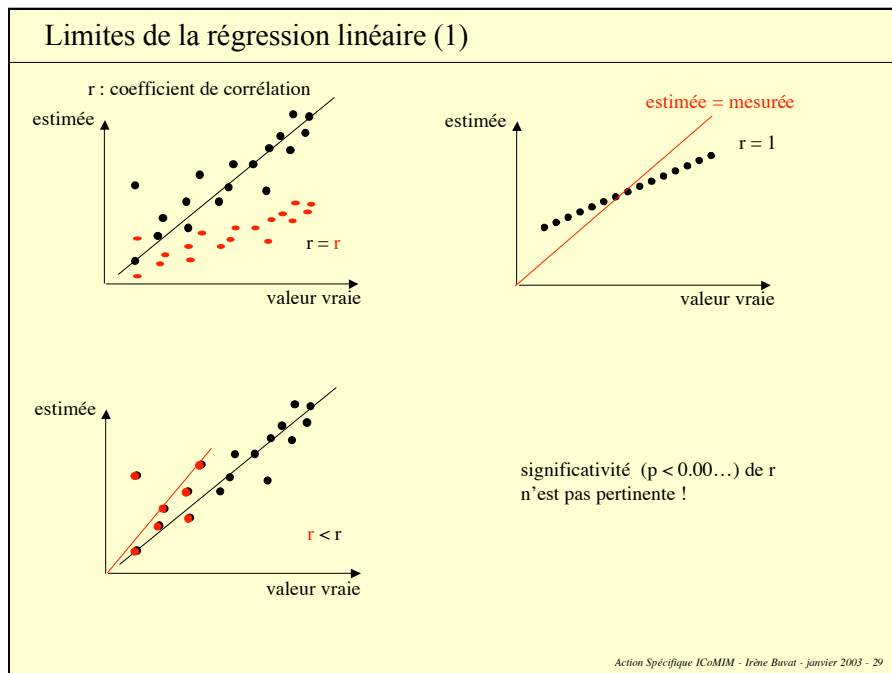
Schématiquement, biais et variabilité peuvent suffire lorsque le biais dépend peu de la valeur du paramètre. En revanche, lorsque le biais dépend sensiblement de la valeur du paramètre, l'évaluation de la méthode nécessite de faire appel à une approche d'évaluation plus complète.

On réalise alors souvent une étude de la régression linéaire entre la valeur estimée et la valeur vraie, pour une large étendue des valeurs vraies potentiellement observables en pratique.

La régression linéaire permet de déterminer :

- si il existe une relation linéaire entre la valeur estimée et la vraie valeur du paramètre. L'existence de cette relation linéaire peut parfois suffire à une interprétation des données.
- si le biais dépend de la vraie valeur du paramètre.

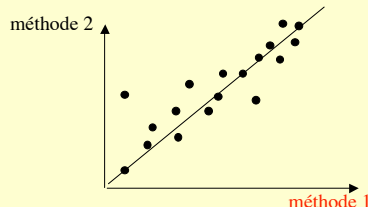
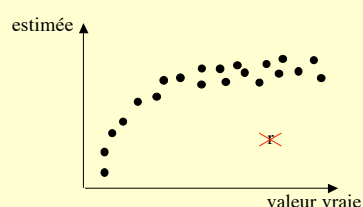
Attention, a priori, la régression linéaire suppose que la référence est connue.



Une simple étude de régression linéaire a cependant ces limites.

- Tout d'abord, on caractérise souvent le résultat d'une régression linéaire par un coefficient de corrélation. On conclut souvent que la méthode est d'autant plus fiable que ce coefficient est élevé. Mais il faut bien avoir conscience que ce coefficient représente le degré de corrélation entre l'estimée et la référence, mais pas l'accord entre ces deux quantités. Autrement dit, on peut parfaitement avoir un  $r = 1$  avec une mesure biaisée.
- La valeur du coefficient de corrélation est insensible au changement d'échelle, alors que le biais l'est bien évidemment.
- La valeur du coefficient de corrélation dépend directement de l'étendue des valeurs de références utilisées pour le calculer. Il est d'autant plus élevé que cette étendue est grande.
- Enfin, certains auteurs basent leurs conclusions sur le degré de significativité de  $r$ . Il est évident que l'estimée est quasiment toujours significativement corrélée au paramètre que l'on cherche à évaluer. La significativité apporte donc très peu d'informations sur la méthode la moins biaisée...

## Limites de la régression linéaire (2)



- \* étude de la corrélation entre les deux méthodes, indépendamment du biais !
- \* peu sensible



L'analyse de la corrélation linéaire entre estimée et valeur vraie est utile pour une interprétation correcte du biais moyen et de sa variabilité.

La caractérisation des performances d'une méthode ou la comparaison des performances de deux méthodes via le coefficient de corrélation est hasardeuse...

Action Spécifique ICoMIM - Irène Buvat - janvier 2003 - 30

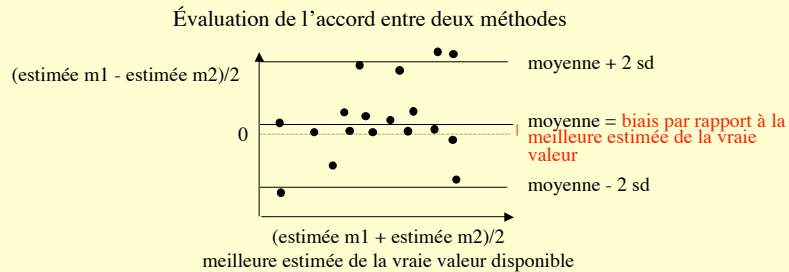
La régression linéaire conduit également à des résultats anormalement pessimistes si la relation entre l'estimée et la valeur vraie n'est pas une relation linéaire. Il est donc indispensable de toujours visualiser cette relation pour juger de la pertinence de la valeur du coefficient de corrélation.

Enfin, souvent, on utilise la régression linéaire pour juger de l'accord entre deux méthodes, lorsqu'aucune d'elles ne peut faire office de gold standard. En fait, nous allons voir que cette méthode est relativement peu sensible, et que lorsqu'on ne dispose pas de gold standard, il existe des approches plus performantes pour mesurer l'accord entre deux méthodes, voire pour estimer la moins biaisée.

En conclusion sur la régression linéaire, l'analyse de la corrélation linéaire entre estimée et valeur vraie est utile pour une interprétation correcte du biais moyen et de sa variabilité.

En revanche, la caractérisation des performances d'une méthode ou la comparaison des performances de deux méthodes via le coefficient de corrélation sont hasardeuses.

## Absence de gold standard : approche de Bland-Altman (1)



Réf : Bland and Altman. *Lancet*, i: 307-10, 1986.



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 !**

Action Spécifique ICoMM - Irène Buvat - janvier 2003 - 31

En l'absence de gold standard, on cherche souvent à caractériser la cohérence des résultats produits par la nouvelle méthode par rapport aux résultats obtenus par la méthode faisant office de standard faute de mieux.

Pour étudier cette cohérence, une approche plus puissante qu'une régression linéaire est l'approche proposée par Bland et Altman. Il s'agit d'une approche simple à mettre en œuvre : il suffit de représenter, pour chaque cas, la différence entre les estimées issues des 2 méthodes en fonction de la moyenne des deux estimées. En l'absence d'informations supplémentaire, cette moyenne représente la meilleure estimée de la valeur vraie du paramètre qui est inconnue.

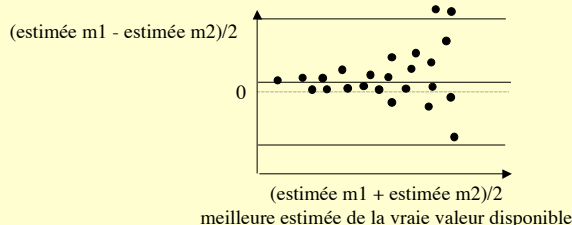
Sur un tel graphe, la moyenne des différences correspond au biais moyen entre les deux méthodes. Si on fait l'hypothèse que les différences suivent une loi normale, 95% des différences seront comprises entre cette valeur moyenne - 2 écart-type des différences, et la valeur moyenne + 2 écart-type des différences. C'est la raison pour laquelle on représente également ces deux lignes sur le graphe.

A quelles questions ce type de graphe permet-il de répondre ?

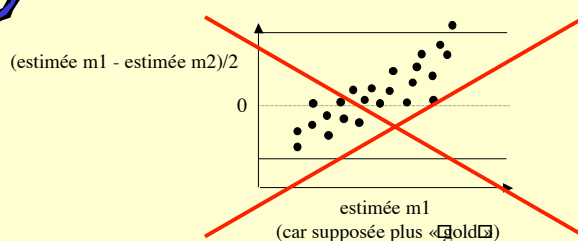
En fonction de l'étendue de cet intervalle, il permet de déterminer si les deux méthodes sont interchangeables (par exemple en se fixant la valeur de différence maximale tolérable pour que les méthodes puissent être utilisées l'une à la place de l'autre).

Cependant, cette approche ne permet en aucun cas de conclure à la supériorité d'une méthode par rapport à une autre, au sens d'une estimation moins biaisée.

## Absence de gold standard : approche de Bland-Altman (2)



Permet de détecter des différences systématiques entre les méthodes



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

Action Spécifique ICoMM - Irène Buvat - janvier 2003 - 32

La représentation de Bland-Altman permet en outre de mettre en évidence des différences entre les méthodes qui dépendraient de la valeur du paramètre à estimer. Par exemple ici, les méthodes produisent des résultats similaires pour de faibles valeurs du paramètres, puis divergents quand la valeur du paramètre augmente.

Des utilisations abusives de la méthode de Bland-Altman ont fait l'objet d'une mise au point par les auteurs.

Parfois, on a tendance à avoir plus confiance dans une méthode que dans une autre, et la représentation de Bland-Altman a été utilisée en représentant la différence des estimées produits par les 2 méthodes en fonction de la valeur estimée par une seule méthode. Bland et Altman montrent que cette représentation est incorrecte, car elle conduit toujours à l'observation d'une tendance, qui n'est pas réelle.



**Régression sans gold standard : détermination de la meilleure méthode**


Hypothèses :

- \* évaluation de la justesse des estimées résultant de la méthode m
- \*  $p_{mi} = a_m p_i + b_m + \epsilon_{mi}$  avec i indiquant le cas
- \*  $p_i$  inconnus
- \*  $\epsilon_m$  suit une loi normale centrée (écart-type  $\sigma_m$ )
- \*  $p_i$  suit une loi de probabilité connue (sans que les paramètres  $r$  de cette loi soient eux même connus, e.g., loi normale)

Méthode :

Détermination des paramètres du modèle qui maximisent la vraisemblance des observables :  $a_m, b_m, \sigma_m, r$

Figure de mérite :  $\sigma_m / a_m$  (le plus faible possible)

 Permet de déterminer la méthode la plus fiable quantitativement en l'absence de gold standard

*Action Spécifique ICoMM - Irène Buvat - janvier 2003 - 33*

Au delà de l'analyse de Bland-Altman, il existe en fait une méthode qui permet d'évaluer la justesse d'une méthode d'estimation en l'absence de gold standard.

Evidemment, il n'y a pas de miracles, donc l'absence de gold standard doit être suppléée par un modèle sur le lien entre les observables et le gold standard inconnu.

Le modèle le plus simple consiste à supposer que l'estimée est une fonction linéaire de la valeur vraie du paramètre,  $p_i$ , qui est elle même inconnue, à une erreur près, qui est supposée gaussienne centrée, mais de variance inconnue.

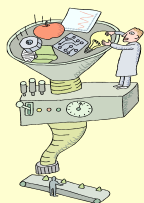
On doit en outre supposée connue le type de loi des  $p_i$ , sans que les paramètres de la loi soient connus. Par exemple, on peut supposer que les vraies valeurs du paramètres suivent une loi normale, de moyenne et d'écart-type inconnus.

La méthode va alors consister à déterminer les paramètres du modèle qui maximisent la vraisemblance des observables. Par exemple, en utilisant un algorithme de type EM, on va obtenir des estimées des paramètres du modèle. Cette estimation revient à faire une régression linéaire par rapport à une variable inconnue. La meilleure méthode sera celle qui conduira à l'erreur la plus faible, c'est à dire au rapport  $\sigma_m/a_m$  le plus faible.

Si on effectue la procédure pour plusieurs méthodes, on peut donc déterminer quelle méthode est a priori la plus fiable.

## Régression sans gold standard : détermination de la meilleure méthode

Caractéristiques de l'approche :



\* 2 méthodes sont suffisantes

\* généralisable à une dépendance non linéaire :

$$p_{mi} = f(\mathbf{p}_i, \mathbf{q}_n) + \mathbf{q}_{mi} \text{ avec } i \text{ indiquant le cas}$$

\* robuste même lorsque l'hypothèse sur la distribution des  $\mathbf{p}_i$  est approximative

\* 25 cas au moins sont nécessaires

Réf: Kupinski et al. *Academic Radiology*, 9: 290-297, 2002  
Hoppin et al. *IEEE Trans Med Imaging*, 21: 441-449, 2002

Action Spécifique ICoMM - Irène Buvat - janvier 2003 - 34

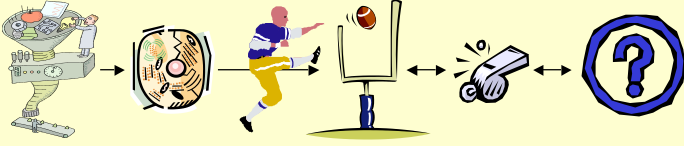
Pour appliquer la méthode avec suffisamment de robustesse, 2 estimées du gold standard par deux méthodes différentes sont nécessaires. On peut également avoir davantage d'estimées, mais le gain en justesse est alors modéré.

Je vous ai présenté la méthode lorsqu'il existe une relation linéaire entre estimée et valeur supposée du gold standard. En fait, la méthode peut tout à fait s'appliquer lorsque cette relation n'est pas linéaire. C'est essentiellement le temps de calcul qui augmente dans ce cas.

Les auteurs ont montré que l'approche reste robuste même lorsque l'hypothèse sur la distribution statistique des  $p_i$  est approximative.

Enfin, concernant le nombre de cas nécessaire, des premiers résultats suggèrent que 25 cas au moins sont nécessaires pour aboutir à une classification correcte des méthodes entre elles.

**Conclusions**



Deux types de travaux d'évaluation :

- tâche de détection
- tâche d'estimation

Pour chaque type, méthodes d'évaluation rigoureuses

- approches type ROC et dérivées
- calcul de biais et variabilité après observation de la dépendance de l'erreur à la valeur vraie

Même en l'absence de gold standard, une évaluation objective rigoureuse est possible

Action Spécifique ICoMIM - Irène Buvat - janvier 2003 - 35

En conclusion, les messages à retenir sont les suivants :

- il existe deux types de travaux d'évaluation, ceux à mettre en œuvre pour les tâches de détection, et ceux concernant les tâches d'estimation.
- Pour chacune de ces tâches, il existe des méthodes rigoureuses d'évaluation objective. Pour les tâches de détection, il s'agit de la famille des méthodes ROC. Pour les tâches d'estimation, il s'agit des calculs de biais et de variabilité, après inspection de la relation entre l'erreur et la valeur vraie du paramètre.
- Enfin, la plupart de ces méthodes connaissent des extensions applicables en l'absence de gold standard. Donc même en l'absence de gold standard, une évaluation objective et rigoureuse est possible.

## Copie de la présentation



<http://www.guillemet.org/irene/equipe4/conferences.html>

Action Spécifique ICoMM - Irène Buvat - janvier 2003 - 36

En conclusion, les messages à retenir sont les suivants :

- il existe deux types de travaux d'évaluation, ceux à mettre en œuvre pour les tâches de détection, et ceux concernant les tâches d'estimation.
- Pour chacune de ces tâches, il existe des méthodes rigoureuses d'évaluation objective. Pour les tâches de détection, il s'agit de la famille des méthodes ROC. Pour les tâches d'estimation, il s'agit des calculs de biais et de variabilité, après inspection de la relation entre l'erreur et la valeur vraie du paramètre.
- Enfin, la plupart de ces méthodes connaissent des extensions applicables en l'absence de gold standard. Donc même en l'absence de gold standard, une évaluation objective et rigoureuse est possible.