UNIVERSITÉ DU DROIT ET DE LA SANTÉ - LILLE 2

FACULTÉ DE MÉDECINE HENRI WAREMBOURG

Année : 2016

THÈSE POUR LE DIPLÔME D'ÉTAT
DE DOCTEUR EN MÉDECINE

Comment mesurer la performance d'un test diagnostique ? Présentation et comparaison d'indicateurs.

Présentée et soutenue publiquement le 16 septembre 2016 à 18 heures
au Pôle Formation
par **Samuel DEGOUL**

JURY

Président :

Monsieur le Professeur Gilles LEBUFFE

Assesseurs :

Monsieur le Professeur Régis BEUSCART
Monsieur le Professeur Alain DUHAMEL
Monsieur le Docteur Jean-Marie RENARD

Directeur de la thèse :

Monsieur le Docteur Emmanuel CHAZARD

La faculté n'entend donner aucune approbation aux opinions émises dans les thèses : celles-ci sont propres à leurs auteurs.

Cette thèse a été préparée au

Centre d'Étude et de Recherche en Informatique Médicale, EA 2694
Faculté de Médecine - Pôle Recherche
1, place de Verdun
59045 Lille cedex
France

☎ +33 (0)3 20 62 69 69
Site `http://cerim.univ-lille2.fr/`

Comment mesurer la performance d'un test diagnostique ? Présentation et comparaison d'indicateurs.

Résumé

Contexte. Ce travail vise à comparer le comportement statistique d'indicateurs synthétiques de performance d'un test diagnostique binaire pour aider le chercheur dans le choix de l'indicateur le mieux adapté pour son étude.

Méthode. Les indicateurs étudiés sont l'aire sous la courbe ROC, le coefficient kappa de Cohen et la proportion de sa valeur maximale, la F-mesure, la concordance observée et le coefficient phi. Au moyen de simulations informatiques basées sur plusieurs valeurs de sensibilité et spécificité du test et de taux de prévalence de l'état à diagnostiquer, les valeurs moyennes et les intervalles de fluctuation de ces indicateurs ont été comparés graphiquement. Différentes situations particulières mais rencontrées en pratique ont été évaluées.

Résultats. Les coefficients kappa et phi se montrent symétriques par rapport aux variations de taux de prévalence, avec une valeur maximale quand les échantillons sont équilibrés. Au contraire, la F-mesure semble anormalement optimiste quand le taux de prévalence est élevé. Elle est de plus le seul des indicateurs étudiés sensible à l'inversion des états « pathologique » et « sain ». La concordance observée surestime systématiquement la performance du test. Il en est de même pour la proportion du kappa maximal, sauf lorsque le taux de prévalence est faible. Enfin, l'aire sous la courbe montre de larges fluctuations de l'intervalle de variation pour des taux de prévalence extrêmes, à l'opposé des autres indicateurs.

Conclusion. Les coefficients kappa et phi semblent être les indicateurs les plus à même de refléter, dans les situations étudiées, la performance d'un test diagnostique en terme d'utilisabilité par les cliniciens. Néanmoins, la valeur d'un indicateur seule ne saurait contenir toute l'information sur cette performance.

Mots clés : test diagnostique, indicateur de performance, aire sous la courbe, kappa de Cohen, F-mesure, coefficient phi

Centre d'Étude et de Recherche en Informatique Médicale, EA 2694
Faculté de Médecine - Pôle Recherche – 1, place de Verdun – 59045 Lille cedex – France

Abréviations

Sigle	Français	Anglais
AF		atrial fibrillation
AFL		atrial flutter
AUC	aire sous la courbe	area under curve
AUC_{norm}	AUC normalisée	normalized AUC
CA	concordance attendue	
CO	concordance observée	
ECG	électrocardiogramme	electrocardiogram
EN	évalué négatif	
EP	évalué positif	
FA	fibrillation auriculaire	
FLA	flutter auriculaire	
F_{meas}		F-measure
F_{mes}	F-mesure	
FN	faux négatif	false negative
FP	faux positif	false positive
κ	coefficient kappa de Cohen	Cohen's kappa coefficient
NPV		negative predicted value
OA		observed agreement
PABAK	kappa ajusté sur la prévalence et le biais	prevalence-adjusted and bias-adjusted kappa
ϕ	coefficient phi	phi coefficient
PN		predicted negative
PP		predicted positive
PPV		positive predicted value
$Prop\ \kappa_{max}$	proportion du kappa maximal	proportion of maximum kappa
RN	réellement négatif	real negative
ROC	« caractéristique du fonctionnement du récepteur »	receiver operating characteristic
RP	réellement positif	real positive
Se	sensibilité	sensitivity
Sp	spécificité	specificity
TN		true negative
TP		true positive
VN	vrai négatif	
VP	vrai positif	
VPN	valeur prédictive négative	
VPP	valeur prédictive positive	

Avant-propos

Ce travail de thèse porte sur un aspect de méthodologie de la recherche sur les tests diagnostiques, à savoir leurs indicateurs de performance. À première vue, ce sujet paraît bien connu des chercheurs et même des cliniciens puisqu'il fait partie du programme des études médicales et est couramment utilisé dans les publications. Ainsi, les notions classiques de « sensibilité » et « spécificité » ou de « valeurs prédictives » ne semblent pas avoir de secret. Il existe aussi des indicateurs synthétiques fournissant une seule valeur de performance « globale » ; ceux-là sont probablement plus connus des chercheurs que des cliniciens.

À regarder de plus près néanmoins, on s'aperçoit que leur utilisation ne prend pas toujours en compte les limites propres à chaque indicateur. De plus, la littérature fournit peu de données permettant la comparaison de ces indicateurs. L'utilisation des indicateurs synthétiques notamment n'est pas homogène, et certains chercheurs préféreront l'aire sous la courbe, d'autres la F-mesure... Le choix d'un indicateur synthétique de performance d'un test diagnostique ne semble donc pas évident.

Notre objectif est avant tout de fournir des éléments aux chercheurs pour orienter ce choix. Ce travail étant destiné à une publication, nous avons organisé le présent document comme suit :

— le premier chapitre (page 1) présente le sujet en décrivant différents indicateurs, leur signification et leur expression mathématique, ainsi que les problématiques qui se posent pour le choix d'un indicateur,

— le deuxième chapitre (page 15) contient l'article, en anglais, soumis à publication,

— le troisième (page 37) reprend et approfondit la discussion de l'article.

On trouvera dans les annexes le script informatique servant de base à nos analyses (page 55) et l'ensemble des résultats graphiques (page 63).

De part une expression simple des résultats sous forme graphique et leur mise en pratique à travers un cas concret d'une étude diagnostique publiée, nous pensons que ce travail intéressera les chercheurs mais aussi les cliniciens soucieux de l'interprétation des indicateurs utilisés dans la littérature.

Table des matières

Résumé **vii**

Remerciements **ix**

Abréviations **xv**

Avant-propos **xvii**

Table des matières **xix**

Liste des tableaux **xxi**

Liste des figures **xxiii**

1 Introduction **1**
1.1 Évaluation d'un test diagnostique 1
1.2 Indicateurs de performance d'un test diagnostique 3
1.2.1 Indicateurs partiels 3
1.2.2 Indicateur synthétiques 5
1.3 Problèmes 11
1.3.1 Utilisabilité 11
1.3.2 Indicateurs partiels *versus* synthétiques 12
1.3.3 Intervalle de confiance 12
1.4 Objectif 12

2 Article **15**
2.1 Background 17
2.1.1 Indicators for the evaluation of diagnostic tests 17
2.1.2 Issues in the choice of an indicator 20
2.1.3 Case study 21
2.1.4 Objectives 21
2.2 Methods 22

2.2.1 Simulations 22
2.2.2 Computations of indicators 23
2.2.3 Variation interval 24
2.2.4 Graphical representations 25
2.3 Results 25
2.4 Discussion 28
2.4.1 Main results 30
2.4.2 Interpretation of the case study 33
2.4.3 Recommendations 34
2.5 Conclusions 36

3 Discussion **37**
3.1 Principaux résultats 37
3.1.1 Concernant l'effet du taux de prévalence 37
3.1.2 Lorsque le taux de prévalence est faible 39
3.1.3 Dans la situation où les diagnostics « positif » et « négatif » sont inversés 39
3.1.4 Concernant l'intervalle de fluctuation 39
3.2 Forces et limites 40
3.3 Interprétation du cas d'étude 41
3.4 Recommandations 43
3.4.1 Différents scénarios 43
3.4.2 Communication des résultats d'une étude diagnostique . . . 44

Conclusion **47**

Bibliographie **49**

A Script **55**

B Ensemble des résultats graphiques **63**

C Informations sur ce travail **67**
C.1 Matériel 67
C.2 Licence 67

Liste des tableaux

1.1 Tableau de contingence 3

2.1 Contingency table 18
2.2 Diagnostic performance indicators of a computerized ECG interpretation reported in Ho *et al.* 22
2.3 Indicators computed from evaluation of a computerized ECG interpretation reported in Ho *et al.* 33

3.1 Indicateurs de performance d'un système d'interprétation automatisée d'électrocardiogramme 41

Liste des figures

1.1 Calcul de l'aire sous la courbe dans le cas d'une courbe ROC avec un seul couple sensibilité / spécificité 7
1.2 Représentation de la relation entre le coefficient kappa de Cohen et la concordance attendue 9

2.1 Mean values of indicators *versus* prevalence for sensitivity = 0.6 and specificity = 0.9 26
2.2 Mean values of indicators *versus* prevalence for sensitivity = 0.9 and specificity = 0.6 27
2.3 Mean values of indicators *versus* log(prevalence) for sensitivity = 0.6 and specificity = 0.9 29
2.4 Mean values of indicators *versus* log(prevalence) for sensitivity = 0.9 and specificity = 0.6 30
2.5 Value of indicators for two opposite situations, "normal" and "inversion" 31

B.1 Variations des indicateurs en fonction du taux de prévalence, pour différents couple de sensibilité (Se) et spécificité (Sp) 64
B.2 Variations des indicateurs en fonction du logarithme du taux de prévalence, pour différents couple de sensibilité (Se) et spécificité (Sp) et en se limitant aux taux de prévalence faibles ($\leq$ 0,1) 65
B.3 Variations des indicateurs selon deux situations opposées, « normale » (taux de prévalence = 0.1, sensibilité = Se, spécificité = Sp) et « inversée » (taux de prévalence = 0.9, sensibilité = Sp et spécificité = Se) 66

Chapitre 1

Introduction

1.1 Évaluation d'un test diagnostique

Un test diagnostique est un examen médical, clinique ou paraclinique, visant à aider le clinicien dans sa démarche diagnostique, qu'il s'agisse d'un diagnostic positif, de gravité ou d'un dépistage. La certitude d'un test diagnostique étant rarement rencontrée, la notion de « probabilité diagnostique » est centrale dans le raisonnement diagnostique [1]. La performance d'un test diagnostique est reflétée par sa capacité de discrimination entre plusieurs classes diagnostiques, plus particulièrement la présence ou l'absence du diagnostic dans le cas d'un test binaire.

L'utilisation d'un test diagnostique pour l'activité de soin suppose une évaluation préalable de sa performance de la même façon qu'un médicament ou un acte thérapeutique. En effet, un test diagnostique peu performant est susceptible d'être associé à de mauvais rapports bénéfice/risque, pour le patient, et coût/efficacité, pour la société. L'évaluation de la performance d'un test diagnostique confronte en général les résultats de ce test à la classification des sujets réalisée par un *gold*

standard. On appelle *gold standard* le meilleur test diagnostique déjà disponible pour la pathologie d'intérêt et utilisable en situation « raisonnable », c'est-à-dire applicable chez des patients en clinique ou en recherche [2]. Ce terme a déjà été soumis à controverses dans la mesure où l'on pourrait croire qu'il est équivalent à « la réalité » [3], alors qu'il ne s'agit que d'un point de référence correspondant aux connaissances dans leur état actuel.

Pour l'intégralité de ce travail, nous étudierons le cas où le résultat du test diagnostique est *binaire*, ne pouvant prendre que deux valeurs exclusives, « positive » et « négative » [4]. Par exemple, ces couples de valeurs peuvent être « présence »/« absence » ou « évolution »/« stabilité » d'une pathologie. Notre choix est justifié par le fait qu'il s'agit d'une situation très fréquente servant de base à la description des différents indicateurs diagnostiques. En effet, même dans le cas où le résultat brut d'un test n'est pas binaire mais quantitatif ou qualitatif ordinal, le choix d'une valeur seuil est souvent effectué pour distinguer deux résultats opposés.

La confrontation des résultats d'un test diagnostique avec les résultats du *gold standard* (abusivement désignés par « réalité » pour la suite) est résumée par un tableau de contingence ou « 2×2 » (tableau 1.1), représentant les effectifs pour chaque catégorie [5]. Les sujets bien classés par le test sont ceux qui sont réellement positifs et évalués positifs par le test, notés vrais positifs (VP), ou ceux qui sont réellement négatifs et évalués négatifs, notés vrais négatifs (VN). À l'opposé, les sujets mal classés par le test sont les faux positifs (FP), réellement négatifs mais évalués positifs, et les faux négatifs (FN), réellement positifs mais évalués négatifs. Les valeurs marginales représentent la division de la population (de taille N) en « réellement positifs » ($RP = VP + FN$) et « réellement négatifs » ($RN = VN + FP$) par le *gold standard*, et en « évalués positifs » ($EP = VP + FP$) et « évalués

négatifs » ($EN = VN + FN$) par le test diagnostique. Le taux de prévalence de la pathologie, c'est-à-dire de la situation « positive », correspond donc à $P = RP/N$.

TABLEAU 1.1 – Tableau de contingence

		réellement	
		positif (RP)	négatif (RN)
évalué	positif (EP)	vrai positif (VP)	faux positif (FP)
	négatif (EN)	faux négatif (FN)	vrai négatif (VN)

1.2 Indicateurs de performance d'un test diagnostique

Les valeurs du tableau de contingence peuvent être combinées pour la construction de différents indicateurs visant à donner une évaluation partielle ou globale de la performance du test diagnostique.

1.2.1 Indicateurs partiels

Les indicateurs partiels de performance d'un test diagnostique ne s'intéressent qu'à l'évaluation de la discrimination d'une des deux classes diagnostiques, positive ou négative.

La sensibilité (notée Se) est la probabilité que le test soit positif chez les sujets réellement positifs, alors que la spécificité (notée Sp) est la probabilité que le test soit négatif chez les sujets réellement négatifs [6] (formules 1.1 et 1.2). Ces indicateurs sont dits « intrinsèques » dans la mesure où ils ne dépendent pas du taux de prévalence de la pathologie dans la population sur laquelle est réalisé le

test [7].

$$Se = \frac{VP}{VP + FN} = \frac{VP}{RP} \tag{1.1}$$

$$Sp = \frac{VN}{VN + FP} = \frac{VN}{RN} \tag{1.2}$$

Néanmoins, la sensibilité et la spécificité apportent peu d'information pour le clinicien en situation réelle de soin. En effet, ces indicateurs sont des probabilités conditionnelles, sachant que le cas est positif ou négatif. Or le clinicien ne sait pas si le patient présente ou non le diagnostic, mais connaît au contraire le résultat du test, qu'il cherche à interpréter [8].

Cette information est fournie par une autre paire d'indicateurs, les valeurs prédictives positives (VPP) et négatives (VPN) correspondant respectivement à la probabilité que le sujet soit réellement positif avec un test positif, et réellement négatif avec un test négatif [9] (formules 1.3 et 1.4). Ils sont qualifiés d'« extrinsèques » puisque leur valeur dépend du taux de prévalence du diagnostic dans la population où est réalisé le test, ce qui limite leur généralisation en pratique car un même test est susceptible d'être utilisé pour des populations différentes ne correspondant pas forcément au contexte de l'étude [10, 11].

$$VPP = \frac{VP}{VP + FP} = \frac{VP}{EP} \tag{1.3}$$

$$VPN = \frac{VN}{VN + FN} = \frac{VN}{EN} \tag{1.4}$$

Enfin, deux autres indicateurs moins connus mais cliniquement pertinents sont

les rapports de vraisemblance, positifs (RVP) et négatifs (RVN). Le premier est défini par le ratio entre la probabilité d'une valeur positive du test chez les sujets réellement positifs et celle chez les sujets réellement négatifs. Le deuxième repose sur le même calcul concernant un résultat négatif du test [12] (formules 1.5 et 1.6). Ces indicateurs, calculés uniquement à partir de la sensibilité et la spécificité, ne dépendent donc pas du taux de prévalence et peuvent à ce titre être qualifiés d'intrinsèques également.

$$RVP = \frac{Se}{1 - Sp} \tag{1.5}$$

$$RVN = \frac{1 - Se}{Sp} \tag{1.6}$$

Ces couples d'indicateurs statistiques appariés permettent l'évaluation séparée des deux facettes de la performance diagnostique, à savoir la conduite à tenir face à un test positif d'une part ou négatif d'autre part. Par contre, il est difficile d'en tirer une idée claire sur la performance globale du test, ce qui est pourtant nécessaire en recherche où l'on souhaite souvent comparer les performances de plusieurs tests diagnostiques.

1.2.2 Indicateur synthétiques

Les indicateurs synthétiques traduisent la performance globale d'un test diagnostique. Si l'on peut trouver leur expression mathématique dans la partie « Méthode » de l'article (section 2.2.2), elles sont reprises ici pour plus de lisibilité.

La courbe ROC (*Receiver Operating Characteristic*) est construite classique-

ment à partir des couples $\{Se,\ Sp\}$ obtenus pour différents seuils diagnostiques retenus pour binariser une variable quantitative en deux résultats, positifs et négatifs [13]. L'aire sous cette courbe (*Area Under Curve*, AUC) correspond à la probabilité de classification correcte des sujets par le test dans le cas d'une variable binaire [14], ou dans le cas d'une variable continue la probabilité qu'un individu présentant le diagnostic, tiré au hasard, ait une valeur plus élevée qu'un individu ne présentant pas le diagnostic (en supposant que le diagnostic soit associé aux valeurs élevées) [13]. L'AUC peut être déterminée facilement dans le cas d'un test d'emblée binaire (sans nécessité de choisir une valeur seuil) [15]. Son calcul (formule 1.7) est illustré figure 1.1. Néanmoins, l'AUC est ici sous-estimée par rapport à un classificateur probabiliste (continu) [16].

$$AUC = Se \times Sp + \frac{Se \times (1 - Sp)}{2} + \frac{Sp \times (1 - Se)}{2} \tag{1.7}$$

Pour faciliter la comparaison avec les autres indicateurs, dont les valeurs se situent entre 0 et 1 dans les conditions habituelles [1], nous avons défini une AUC « normalisée » pour avoir une répartition sur la même gamme de valeurs (formule 1.8).

$$AUC_{norm} = 2 \times AUC - 1 \tag{1.8}$$

La concordance observée (CO) traduit la précision « brute », non ajustée. Elle correspond à la proportion de sujets bien classés, qu'ils soient positifs ou négatifs (formule 1.9). Néanmoins, une partie de cette concordance est liée au ha-

1. Nous entendons par « conditions habituelles » un test diagnostique assez bien fait pour ne pas classer les individus plus mal que ne le ferait une répartition aléatoire.

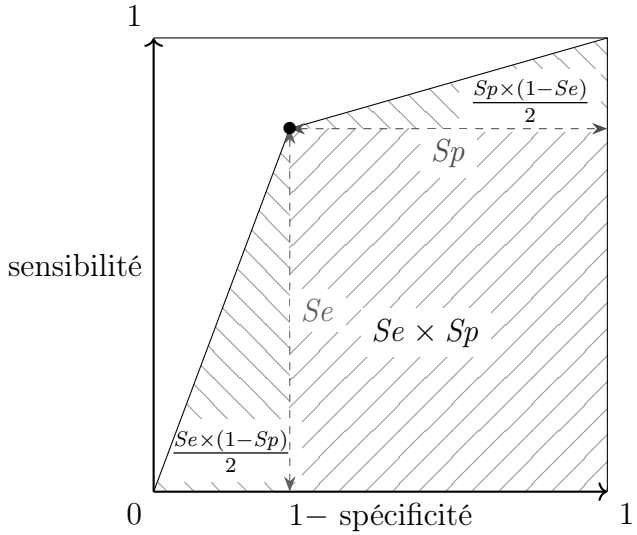


FIGURE 1.1 – Calcul de l'aire sous la courbe (AUC) dans le cas d'une courbe ROC avec un seul couple sensibilité (Se) / spécificité (Sp). L'AUC est divisée en trois surfaces dont les tailles sont données par les formules.

sard, s'expliquant par les probabilités jointes des valeurs marginales du tableau de contingence [17]. Ainsi, si le résultat du test était entièrement aléatoire (avec un probabilité p de 0,5 par exemple), pour un taux de prévalence donné P, le nombre de VP serait de $0,5 \times P \times N$ et le nombre de VN serait de $0,5 \times (1-P) \times N$ d'où une concordance de 0,5 alors que la performance réelle du test est nulle.

$$CO = \frac{VP + VN}{VP + FP + FN + VN} \tag{1.9}$$

La coefficient kappa de Cohen (κ) mesure une concordance ajustée sur cette concordance qui serait obtenue par hasard. Il a été développé pour évaluer la concordance entre deux juges [17]. Dans notre cas, un juge est représenté par le *gold standard*. Cet indicateur est une des implémentations de mesure de concordance ajustée sur le hasard. Ici, le hasard est conditionnel aux valeurs marginales. Si, comme nous l'avons vu dans le paragraphe précédent, cette concordance « par

hasard » est de 0,5 dans le cas où les marges sont équilibrés (c'est-à-dire autant de patients positifs que négatifs), ce n'est pas le cas si le nombre de sujets positifs et négatifs est très différent. La concordance attendue (CA), « par hasard », est déterminée à partir des valeurs attendues (A) pour les vrais positifs $VP_A = (RP \times EP)/N$ et les vrais négatifs $VN_A = (RN \times EN)/N$. La formule 1.10 détaille le calcul du κ et la figure 1.2 illustre sa signification.

$$CA = \frac{VP_A + VN_A}{N}$$
$$\kappa = \frac{CO - CA}{1 - CA} = \frac{2(VP \times VN - FP \times FN)}{(VP + FN)(FN + VN) + (FP + VN)(VP + FP)} \quad (1.10)$$

Sa détermination repose sur l'hypothèse d'indépendance des évaluateurs et l'indépendance des sujets évalués [7, 17]. Le κ a été décrié devant des comportements d'apparence paradoxale dans deux situations :

— en cas de déséquilibre entre les effectifs positifs et négatifs (effet prévalence), conduisant à un κ apparemment trop bas,

— en cas de différences importantes de distributions marginales entre les évaluateurs (effet biais), responsable d'une surestimation apparente de la concordance.

Néanmoins, même si des méthodes de corrections de ces effets ont été développées, tel le PABAK (*prevalence-adjusted and bias-adjusted kappa*), les valeurs apparemment trop basses du κ semblent bien refléter une réelle mauvaise concordance, et l'utilisation du PABAK n'est pas conseillée du fait d'une inflation du κ dans le cas de la correction de l'effet prévalence, et une négativation dans le cas de l'effet biais [19].

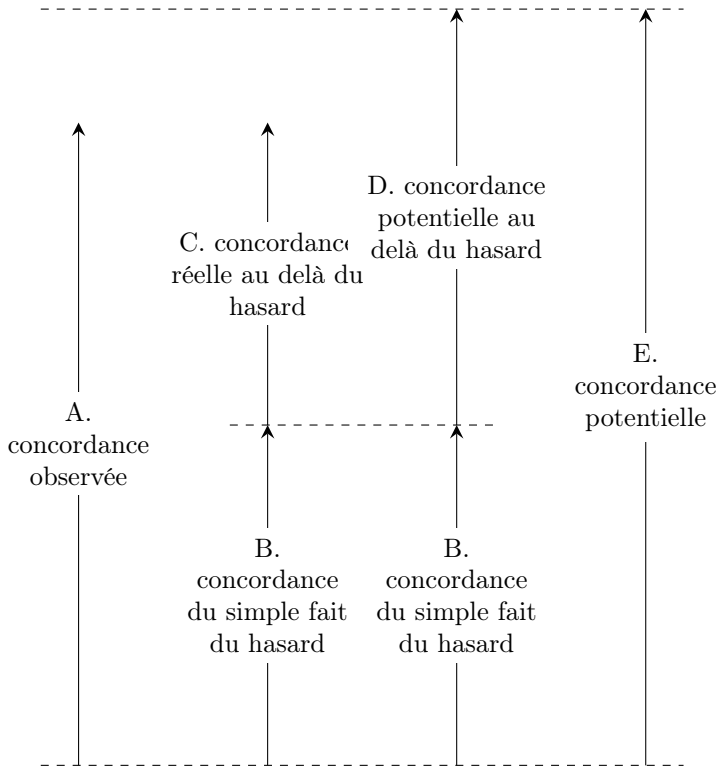


FIGURE 1.2 – Représentation de la relation entre le coefficient kappa de Cohen et la concordance attendue (due au hasard). $\kappa = C/D$. Tiré de Sim et Wright [18]

Un indicateur directement dérivé du κ est la proportion du kappa maximal (*Prop* κ_{max}). En effet, en cas de déséquilibre des valeurs marginales, la valeur du κ ne peut pas toujours atteindre 1 [18]. On peut donc déterminer la valeur maximale théorique, conditionnelle aux marges, et y rapporter la valeur observée du κ (formule 1.11).

$$\kappa_{max} = \frac{\frac{CO^2}{2}}{1 - CO + \frac{CO^2}{2}} = \frac{CO^2}{(1 - CO)^2 + 1}$$
$$Prop\ \kappa_{max} = \frac{\kappa}{\kappa_{max}} \tag{1.11}$$

La F-mesure (F_{mes}) est un autre indicateur, correspondant à la moyenne harmonique de la *Se* et de la *VPP* [20] (formule 1.12). On appelle score F1 le cas particulier où le facteur de pondération est de 1, cas le plus courant [20] et que nous retiendrons par la suite. Une caractéristique de cet indicateur est de pouvoir se passer de la connaissance du nombre de vrais négatifs, propriété intéressante dans le champs de recherche de l'extraction d'informations et la fouille de texte d'où est né cet outil (van Rijsbergen en 1975). En effet, il est parfois impossible de déterminer le nombre, potentiellement infini, de documents ne contenant pas l'information recherchée.

$$F_{mes} = \frac{2 \times Se \times VPP}{Se + VPP} = \frac{2 \times VP}{2 \times VP + FP + FN} \tag{1.12}$$

Enfin, le coefficient de corrélation phi (ϕ), ou coefficient de Yule 1, correspond à l'application du coefficient de corrélation de Pearson à une variable binaire [21] et traduit l'association entre ces deux variables. Son calcul est détaillé formule 1.13.

$$\phi = \frac{VP \times VN - FP \times FN}{(VP + FP)(VP + FN)(VN + FP)(VN + FN)} \tag{1.13}$$

Cette liste n'est pas exhaustive et il existe en fait un nombre très important de coefficients d'association pour les tables 2×2. Nous avons ici choisi de nous intéresser aux indicateurs les plus connus mais aussi basés sur des principes de calculs bien distincts.

Nous n'envisageons pas ici la statistique du χ^2 qui est est un marqueur d'association entre deux variables qualitatives mais pas un indicateur de concordance. Ainsi, une association significative, c'est-à-dire non expliquée par le hasard, peut refléter un accord mais aussi un désaccord important entre les deux évaluateurs

[17]. Un juge étant le *gold standard*, si le deuxième évaluateur est un vrai juge, même médiocre, et non un tirage au sort, l'association statistique sera significative. D'autre part, le χ^2 ne fournit pas de taille d'effet comme les autres indicateurs mais seulement une information sur la significativité.

1.3 Problèmes

1.3.1 Utilisabilité

Devant l'existence de nombreux indicateurs – et nous n'avons présenté ci-dessus que les principaux – le choix de l'un d'entre eux n'est pas aisé pour le chercheur qui veut évaluer la performance d'un test diagnostique.

Le couple {sensibilité, spécificité} est souvent utilisé et bien connu parmi les cliniciens. Néanmoins, comme expliqué plus haut, l'information apportée par ces indicateurs concerne les caractéristiques intrinsèques du test, sans prendre en compte le contexte d'application. Ainsi, un test doté d'une spécificité considérée comme « bonne » pourra montrer des performances médiocres en cas de pathologie rare, avec de nombreux faux positifs.

Les indicateurs extrinsèques, qui peuvent prendre en compte le contexte d'application du test, paraissent plus utilisables en pratique. Néanmoins, leur application est plus délicate puisqu'il faut alors prendre en compte ce contexte, c'est-à-dire le taux de prévalence.

1.3.2 Indicateurs partiels *versus* synthétiques

L'utilisation des indicateurs appariés n'est pas évidente pour évaluer la performance d'un test, dans la mesure où ils nécessitent la synthèse des deux informations complémentaires qu'ils apportent. Il est ainsi difficile de comparer la fiabilité de deux tests concurrents, surtout si chacun domine l'autre sur un seul indicateur.

Les indicateurs synthétiques répondent à cette nécessité de disposer d'une information globale, permettant de comparer aisément plusieurs tests. Il est par contre plus difficile de se faire une représentation de ce qu'ils signifient.

1.3.3 Intervalle de confiance

Toute estimation statistique déterminée à partir d'un échantillon, comme c'est le cas pour une étude d'évaluation d'un test diagnostique, est entachée d'incertitude justifiant l'utilisation d'intervalle de confiance. Si ceci est bien connu en épidémiologie et en recherche clinique, l'utilisation d'intervalle de confiance est moins courante dans l'évaluation des tests diagnostiques, même si elle est recommandée [5].

De plus, parmi les multiples indicateurs disponibles, on ne dispose pas de comparaison de leur précision d'estimation, information qui pourrait faire choisir l'un par rapport à l'autre.

1.4 Objectif

L'objectif de ce travail est de fournir au chercheur des outils pour l'aider à discuter le choix d'un indicateur synthétique de performance dans le cadre d'une

étude d'évaluation d'un test diagnostique. Il s'agira d'une approche didactique basée sur des outils graphiques. Nous étudierons également des contextes particuliers rencontrés en médecine, à savoir le cas où la pathologie est rare et celui où les situations « normale » et « pathologique » sont inversées.

La méthode choisie pour ce travail sera basée sur une comparaison graphique de la valeur et de l'intervalle de fluctuation (ou de variation) des différents indicateurs synthétiques présentés plus haut, déterminés par simulation informatique pour une taille d'échantillon choisie arbitrairement.

Nous illustrerons nos propos par un exemple réel provenant de la cardiologie, plus précisément l'analyse automatisée d'électrocardiogramme (ECG). En effet, l'interprétation du tracé que fournit un appareil ECG fait appel à plusieurs tests diagnostiques pour déterminer la présence ou l'absence d'un trouble de conduction, un trouble du rythme... Plusieurs systèmes ont ainsi été développés, qu'il est nécessaire d'évaluer. À ce titre, *Ho et al.* ont publié les résultats d'une étude diagnostique ([22]) évaluant un système d'interprétation automatique d'ECG sur une base de données existante d'enregistrements électrocardiographiques, travail que nous prendrons comme cas d'étude.

Chapitre 2

Article

Ce chapitre présente l'article lié à ce travail tel qu'il a été soumis au journal *BMC Medical Research Methodology*[1] le 12 juillet 2016.

Pour assurer une mise en page cohérente avec le reste du document, les numéros des sections, tableaux et figures ont été adaptés, ainsi que les références vers ce qui a été adressé en *supplementary material* avec la soumission. Les références bibliographiques, dont la plupart sont communes avec le reste du document, se trouvent dans la liste des références à la fin du document.

1. `http://bmcmedresmethodol.biomedcentral.com/`

Which indicator should be used to evaluate the performance of a diagnostic test?

Samuel Degoul, Grégoire Ficheur, Émilie Baro, Régis Beuscart and Emmanuel Chazard

Abstract

Background. This study aims to compare the behavior of some aggregated indicators of diagnostic test performance in order to provide advice on the choice of indicator. An illustration of computerized interpretation of electrocardiogram is presented.

Methods. The bench test is a binary diagnostic test. Multiple simulations are performed based on predefined sets of sensitivity and specificity of the test, and prevalence of the disease. Indicators of interest are the area under curve, the observed agreement, the Cohen's kappa coefficient and the proportion of its maximum value, the F-measure, and the phi coefficient. Mean values and 95% variation intervals of these indicators are graphically compared in different scenarios.

Results. The kappa and the phi coefficients are roughly symmetrical along all possible values of prevalence, with a maximum for a balanced sample. On the contrary, the F-measure is biased when the prevalence is high. Moreover, the F-measure is the only indicator sensitive to the inversion of ill and healthy state. The observed agreement and the proportion of maximum kappa seem to overestimate the performance of the test. Finally, contrary to other indicators whose estimation is

precise, the area under curve has a large variation interval for extreme prevalence.

Conclusions. The kappa and the phi coefficient seem to be the most reliable indicators in several situations to assess the performance of a diagnostic test. Nevertheless, only one indicator may not be sufficient to encompass the entire information of the performance of the test.

2.1 Background

Within the framework of computerized interpretation of electrocardiogram (ECG), each message provided by the device (e.g., myocardial infarction or atrial fibrillation) is the result of an independent diagnostic test. This test has a binary outcome insofar as it is designed to discriminate between two states of health. Throughout this paper, we consider a two-alternative forced-choice test, which concerns most of diagnostic tests [4].

2.1.1 Indicators for the evaluation of diagnostic tests

Diagnostic tests are used to aid in the diagnosis or detection of diseases [1].

For a daily use, the clinical utility of a diagnostic test must be evaluated to prevent the use of a test with a low cost-effectiveness ratio or an unfavorable risk-benefit ratio.

Diagnostic performance studies commonly compare the test to a "gold standard" test, which is a diagnostic test that is the best available under reasonable conditions [2]. The classification accuracy of the test under evaluation is usually summarized in a 2-by-2 contingency table [5]. Table 2.1 shows a contingency table

for a dichotomous classification, with "positive" and "negative" outcomes. Marginal totals are count of subjects for each class: real positives (RP) and real negatives (RN), and predicted positives (PP) and predicted negatives (PN) are determined by the gold standard and the test under evaluation, respectively. Letting N represent the total number of subjects, N is equal to $RP + RN = PP + PN$ and the prevalence rate of positive cases is RP/N. Cell values are raw count of the number of times each predicted class is associated with each real class. True positives (TP) and true negatives (TN) refer to the number of correct classifications for RP and RN, respectively. Conversely, false positives (FP) and false negatives (FN) refer to incorrect classifications for RN and RP, respectively.

Table 2.1 – Contingency table

		real	
		positive (RP)	negative (RN)
predicted	positive (PP)	$TP^{\S}$	FP^{**}
	negative (PN)	FN^{*}	TN^{+}

*false negative, **false positive, +true negative, §true positive

Various indicators are usually computed to assess the diagnostic performance of the test.

Well-known indicators are sensitivity (or recall) and specificity [6]. Sensitivity (Se) is the proportion of RP cases which are correctly predicted positive, whereas specificity (Sp) is the proportion of RN cases which are correctly predicted negative. These indicators are prevalence-independent. Mirror indicators are positive predictive value (PPV), also known as precision, and negative predictive value (NPV) [9], respectively defined by the proportion of PP which are actually positive and the proportion of PN which are actually negative. Unlike Se and Sp,

PPV and NPV depend on prevalence which is bound to the study design and the conditions of experimentation. Besides, positive and negative likelihood ratios correspond to the ratio of the probability of a positive or negative result in RP subjects to the probability in RN subjects [12]. Because they do not depend on prevalence and they adapt for varying prior probabilities, these alternative statistics are considered more useful clinically. Nevertheless, each of these indicators only partially assess the diagnostic performance because they only focus on positive or negative cases.

Many aggregated indicators have been developed to encompass both positive and negative cases so that they can reflect the global ability of the test to discriminate the diagnostic groups.

Receiver operating characteristic (ROC) curve can be drawn by plotting Se against $1-Sp$ [14]. The area under the curve (AUC) corresponds to the probability of correct classification of subjects.

Observed agreement (OA) corresponds to the raw accuracy of the test: $(TP + TN)/N$. Nevertheless, this indicator is biased because it may be at least partly explained by chance alone.

Cohen's Kappa coefficient (κ) has been developed to assess the agreement between two independent raters for qualitative variables [7]. In the particular case of a diagnostic test, one of the raters is the gold standard. Kappa calculation is based on correction of observed agreement beyond chance. Its assumptions are independence of raters and independence of subjects. Paradoxical variations of κ have been described which lead to low value of κ contrasting with high observed agreement in case of unbalanced sample (prevalence effect) or large differences in marginal totals reflecting differences in evaluation between raters (bias effect). A

prevalence-adjusted and bias-adjusted κ (PABAK) has been developed. However, these apparent paradoxes can be considered to reflect an actual poor agreement. PABAK seems to be biased and its use is not recommended [19]. Another correction that has been proposed is the proportion of maximum value of κ ($Prop\ \kappa_{max}$). Indeed, for a given situation, κ cannot exceed a maximum value, which can be computed [23].

F-measure (F_{meas}), also called F1 score when weighting is unity, is a weighted harmonic mean of Se and PPV [24, 20]. It is used in information retrieval studies which usually lack a well-defined number of real negative cases [25].

Phi correlation coefficient (ϕ), also called coefficient of Yule 1, is another association coefficient for 2-by-2 tables and acts as the application of the Pearson correlation coefficient to a binary variable [21].

2.1.2 Issues in the choice of an indicator

It is not so straightforward to choose the best indicator to reflect the diagnostic performance of a diagnostic test. A distinction between accuracy and applicability in real population has to be made. Accuracy parameters of a diagnostic test, such as Se and Sp, reflect the intrinsic quality of the test, i.e., its ability to discriminate diagnostic classes under study conditions. Nevertheless, this accuracy does not go hand in hand with the quality of discrimination under "real life" conditions, which concerns the applicability of the test. Good Se and Sp can thus be associated with poor performance in clinical practice [8]. In fact, applicability is one of the main criteria in the QUADAS-2 diagnostic test evaluation tool [5, 26].

Furthermore, the use of combined statistics (partial indicators) leads to a dif-

ficult interpretation of the global goodness of the test under evaluation, while an aggregated indicator can be useful for comparison of several tests.

Finally, variation intervals of these indicators are not widely known and have not yet been compared.

2.1.3 Case study

Our analysis is built on a real example of computerized electrocardiogram (ECG) interpretation.

Many detection algorithms have been developed and compared. In particular, Ho *et al.* [22] developed a remote monitoring system with computerized ECG interpretation which was evaluated against 213,420 heartbeats from a clinical ECG database (gold standard). We focus on three diagnostic tests compared in this study, aiming for diagnosing sinus rhythm and two types of cardiac arrhythmia, i.e., atrial fibrillation (AF) and atrial flutter (AFL). Results presented in this study are summarized in table 2.2 and rise difficulties in interpretation: at first glance, one might think that the system performs better for AFL detection with an accuracy (OA) of 0.919. However, when comparing to AF detection, we could draw a different conclusion because the Se is lower for AFL detection than for AF detection, whereas their Sp are close.

2.1.4 Objectives

The aim of this work is to present a didactic approach for providing advice and graphical tools to help researchers in the choice of an aggregated diagnostic to assess the diagnostic performance of a test, especially in the field of computerized

Table 2.2 – Diagnostic performance indicators of a computerized ECG interpretation reported in Ho *et al.* [22]

cardiac rhythm	$TP^{\S\S}$	FN^{*}	FP^{**}	$TN^{\S}$	Se^{+}	Sp^{++}	OA^{***}
sinus	47036	85575	1824	52804	0.355	0.967	0.533
atrial fibrillation	17935	1413	20357	147534	0.927	0.879	0.884
atrial flutter	4391	2667	12530	167651	0.622	0.934	0.919

*false negative, **false positive, ***observed agreement, $^{+}$sensitivity, $^{++}$specificity, §true negative, §§true positive

ECG interpretation.

To this end, a simulation-based study is carried out to compare these indicators. The simulation method allows the use of a single method for all indicators with the aim of a simple comparison of their values and variation intervals.

2.2 Methods

2.2.1 Simulations

Simulations are based on an evaluation of a diagnostic test *versus* a gold standard test in its ability to detect a binary outcome with positive and negative cases coded 1 and 0, respectively.

A population of $N = 10^4$ individuals submitted to the test is considered. Contingency table data are simulated based on predefined population characteristics of the diagnostic test: prevalence of positive cases (P), sensitivity (Se), and specificity (Sp) of the test. After allocating the value 1 to the results of the gold standard for $N \times P$ individuals (other individuals having 0), positive result of the diagnostic test (value = 1) is allocated statistically, with a probability equal to Se

for individuals positive with the gold standard, and a probability equal to $1 - Sp$ for individuals negative with the gold standard. A random number generator and a uniform law are used for the simulations.

The sets of values of population characteristics are {0.5, 0.6, 0.7, 0.8, 0.9} for Se and Sp and {10^{-3}, $10^{-2.667}$, $10^{-2.333}$, 10^{-2}, $10^{-1.667}$, $10^{-1.333}$, 10^{-1}, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 0.99} for prevalence.

Simulations are repeated 10^3 times for each combination of these three parameters.

2.2.2 Computations of indicators

AUC calculation (formula 2.1) is simple with only one set of Se and Sp [15] and is illustrated figure 1.1. It should be noted that in the case of a test with continuous response, this computation underestimates the AUC. However, the present study is based on a test with a binary response.

$$AUC = Se \times Sp + \frac{Se \times (1 - Sp)}{2} + \frac{Sp \times (1 - Se)}{2} \tag{2.1}$$

Normalized AUC, $AUC_{norm} = 2 \times AUC - 1$, is defined for easier comparison with other indicators because of a similar range of values (from 0 to 1).

Observed agreement is defined by the formula 2.2.

$$OA = \frac{TP + TN}{TP + FP + FN + TN} \tag{2.2}$$

For κ computation, the observed agreement is defined above. For TP cell, the expected value "by chance" is $E_{TP} = (RP \times PP)/N$, and for TN cell, $E_{TN} = (RN \times PN)/N$. Thus, the overall agreement expected by chance is $EA = (E_{TP} + E_{TN})/N$.

Kappa formulation (formula 2.3) is given by Hripcsak and Heitjan [27].

$$\kappa = \frac{OA - EA}{1 - EA} = \frac{2(TP \times TN - FP \times FN)}{(TP + FN)(FN + TN) + (FP + TN)(TP + FP)} \tag{2.3}$$

Formula 2.4 shows computation of *Prop* κ_{max} from maximum value of κ (κ_{max}), which uses OA.

$$\kappa_{max} = \frac{\frac{OA^2}{2}}{1 - OA + \frac{OA^2}{2}} = \frac{OA^2}{(1 - OA)^2 + 1}$$
$$Prop\ \kappa_{max} = \frac{\kappa}{\kappa_{max}} \tag{2.4}$$

For F_{meas}, considering a balanced effect of Se and PPV (F1 score), i.e., a weight coefficient $\beta = 1$, its computation is given by formula 2.5:

$$F_{meas} = \frac{2 \times Se \times PPV}{Se + PPV} = \frac{2 \times TP}{2 \times TP + FP + FN} \tag{2.5}$$

Finally, ϕ is computed as shown in formula 2.6 [28].

$$\phi = \frac{TP \times TN - FP \times FN}{(TP + FP)(TP + FN)(TN + FP)(TN + FN)} \tag{2.6}$$

2.2.3 Variation interval

The above-mentioned indicators are estimated for each sample, and their empirical distribution is used to estimate the 95% variation interval, between the 2.5^{th} and 97.5^{th} percentiles.

2.2.4 Graphical representations

Mean values and 95% variation interval of each indicator are plotted against P for each combination of Se and Sp, and then against $Log(P)$ for low prevalence (P lower than or equal to 0.1).

A scenario of "inversion" is considered. It shows the consequence of the use of the test to detect a healthy state instead of the disease (e.g., normal sinus cardiac rhythm). In that case, P becomes $1-P$, Se becomes Sp and Sp becomes Se. This situation is illustrated by plotting mean values and variation interval of indicators against both opposite prevalences.

In this paper, all charts are shown for the couple $\{Se = 0.6, Sp = 0.9\}$ which roughly corresponds to the case of AFL diagnostic performance from the case study (table 2.2), together with the opposed situation $\{Se = 0.9, Sp = 0.6\}$.

All analyses and graphical representations are performed with R software version 3.2.2 for Linux [29]. The script is available in appendix A.

2.3 Results

Figures 2.1 and 2.2 show the variations of the indicators with prevalence for two opposite combinations of Se and Sp. By construction, mean value of AUC_{norm} is constant because of fixed theoretical Se and Sp. Three indicators, κ, *Prop* κ_{max} and ϕ, are roughly symmetrical, with a maximum for prevalence around 0.5 and minimum for extreme values of prevalence. The range of values of these indicators increases with Se and Sp. The κ and ϕ curves are skewed toward low prevalences when Se is lower than Sp or high prevalences when Se is higher than Sp. The *Prop* κ_{max} curve is skewed in the opposite direction.

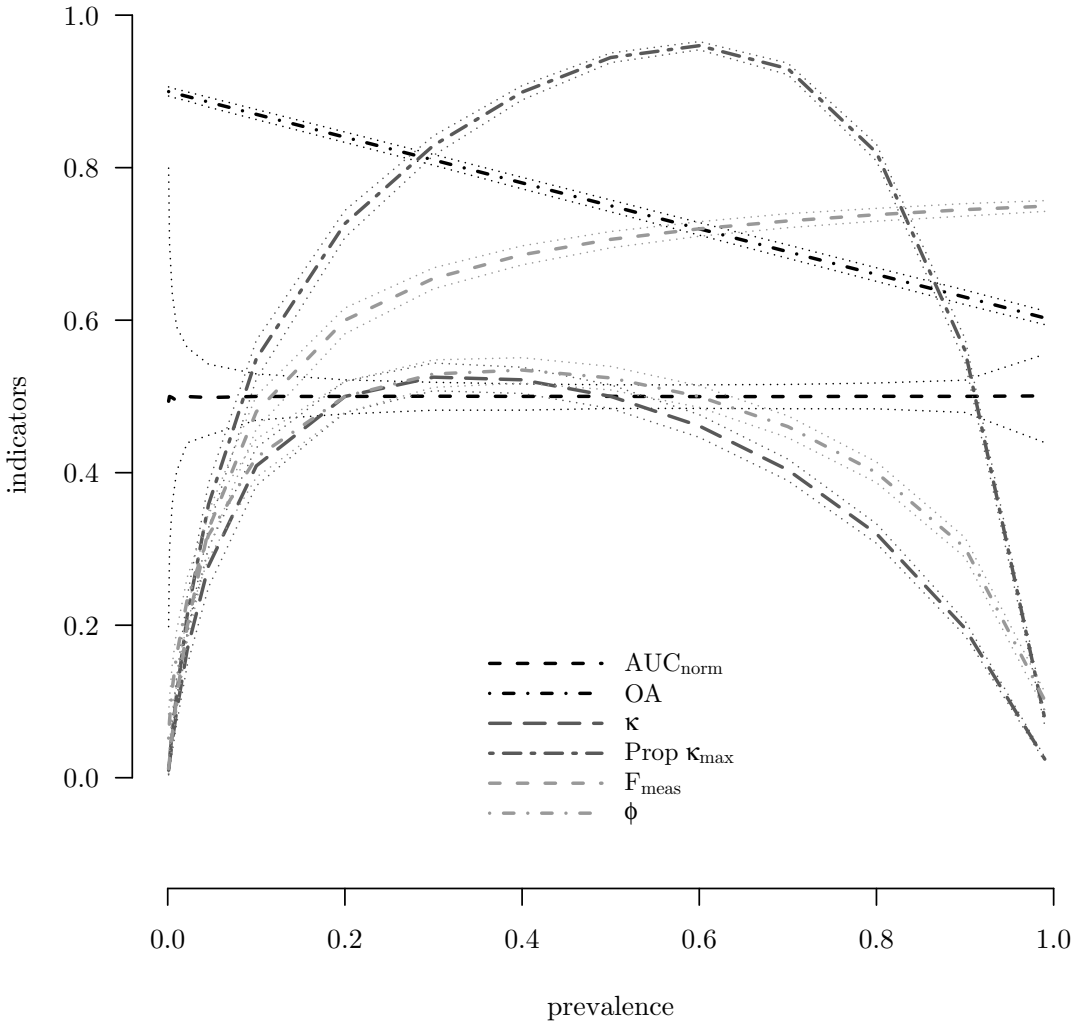


Figure 2.1 – Mean values of indicators *versus* prevalence for sensitivity = 0.6 and specificity = 0.9. AUC_{norm}: normalized area under curve, κ: Cohen's kappa, F_{meas}: F-measure, *OA*: observed agreement, ϕ: phi coefficient, *Prop* κ_{max}: proportion of maximum kappa (thin dotted lines represent variation intervals)

Conversely, other indicators are asymmetric: F_{meas} is consistently maximal for highest prevalences and minimal for low prevalences. For *OA*, slope direction depends on the predominance of *Se* over *Sp*: constant when *Se* is equal to *Sp* (appendix B figure B.1), increasing with prevalence when *Se* is higher than *Sp*

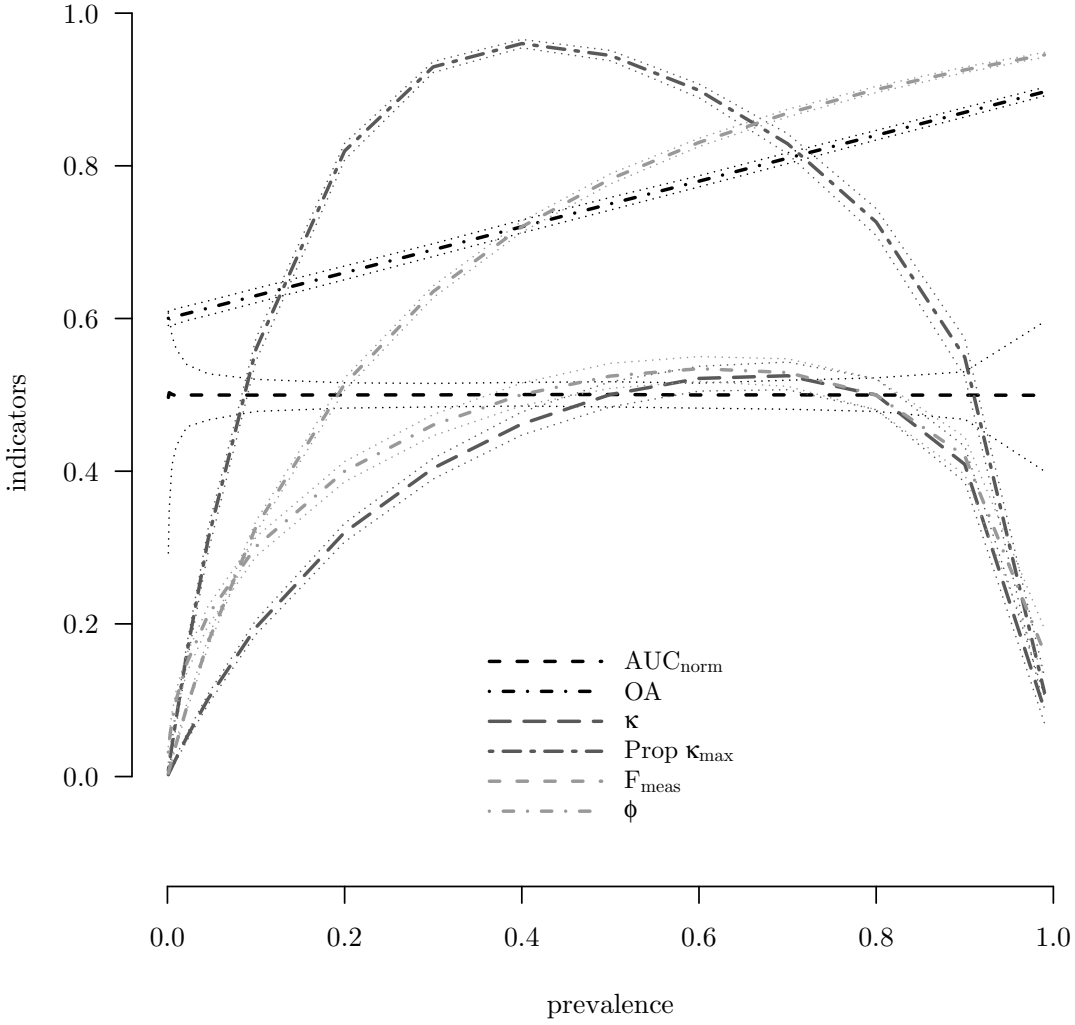


Figure 2.2 – Mean values of indicators *versus* prevalence for sensitivity = 0.9 and specificity = 0.6. AUC_{norm}: normalized area under curve, κ: Cohen's kappa, F_{meas}: F-measure, OA: observed agreement, ϕ: phi coefficient, *Prop* κ_{max}: proportion of maximum kappa (thin dotted lines represent variation intervals)

and decreasing when Se is lower than Sp.

Regarding comparison of values between indicators, it can be observed that κ and ϕ curves are rarely superior to AUC_{norm} curve, and reach it only under optimal condition, i.e., P around 0.5. On the contrary, OA stands over the AUC_{norm} curve

in the range of values between *Se* and *Sp*. *Prop* κ_{max} and F_{meas} rapidly go over AUC_{norm} curve when getting away from extreme prevalences for *Prop* κ_{max}, and only with high prevalences for F_{meas}. The discrepancy between F_{meas} and κ or ϕ is even greater when *Se* is much higher than *Sp*. This is explained by the fact that F_{meas} takes much more into account *Se* than *Sp* whereas both other indicators encompass equally *Se* and *Sp*.

For most indicators, variation interval is tight and does not depend on prevalence. An irregular variation interval is only observed for AUC_{norm} with an important enlargement for both low and high prevalences.

Focusing on prevalences lower than or equal to 0.1, figures 2.3 and 2.4 show that value of κ, *Prop* κ_{max}, ϕ, and F_{meas} are close, even if *Prop* κ_{max} and F_{meas} tend to increase more steeply than κ and ϕ. AUC_{norm} is associated with a large variation interval contrasting with other indicators (see appendix B figure B.2).

In the case of the detection of "normal" state instead of "pathological" state (figure 2.5), only F_{meas} is sensitive to the situation and seems to be overvalued. Other indicators remain stable through condition change (see appendix B figure B.3).

2.4 Discussion

In this work, we used simulation as a common benchmark which allows straightforward comparison of indicators behavior and leads to a simple but robust estimation of their variation interval.

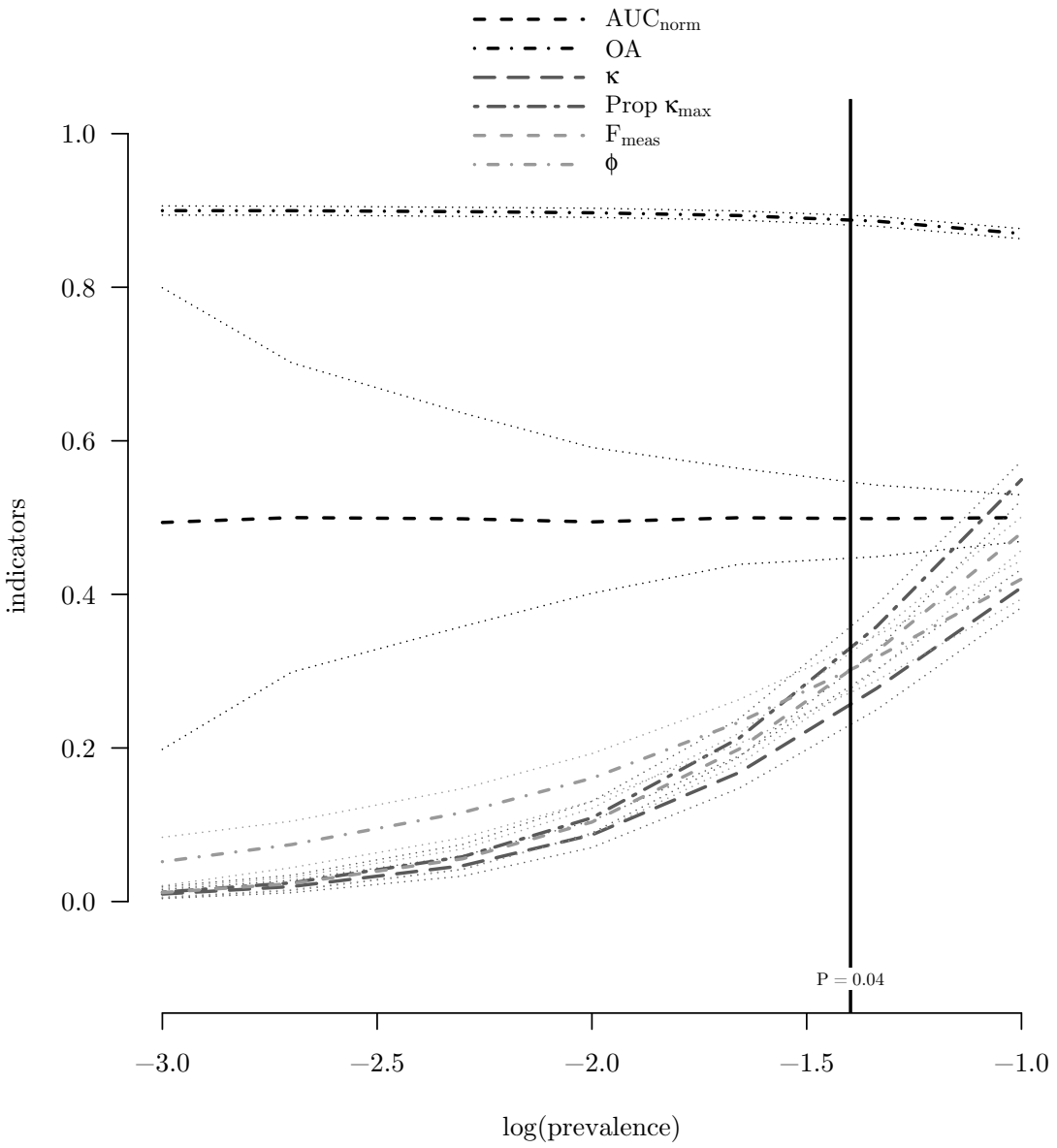


Figure 2.3 – Mean values of indicators *versus* log(prevalence) for sensitivity = 0.6 and specificity = 0.9. The solid vertical line represents the case of AFL diagnosis in Ho *et al.* [22] with a prevalence (P) of 0.04. AUC_{norm}: normalized area under curve, κ: Cohen's kappa, F_{meas}: F-measure, OA: observed agreement, ϕ: phi coefficient, *Prop* κ_{max}: proportion of maximum kappa (thin dotted lines represent variation intervals)

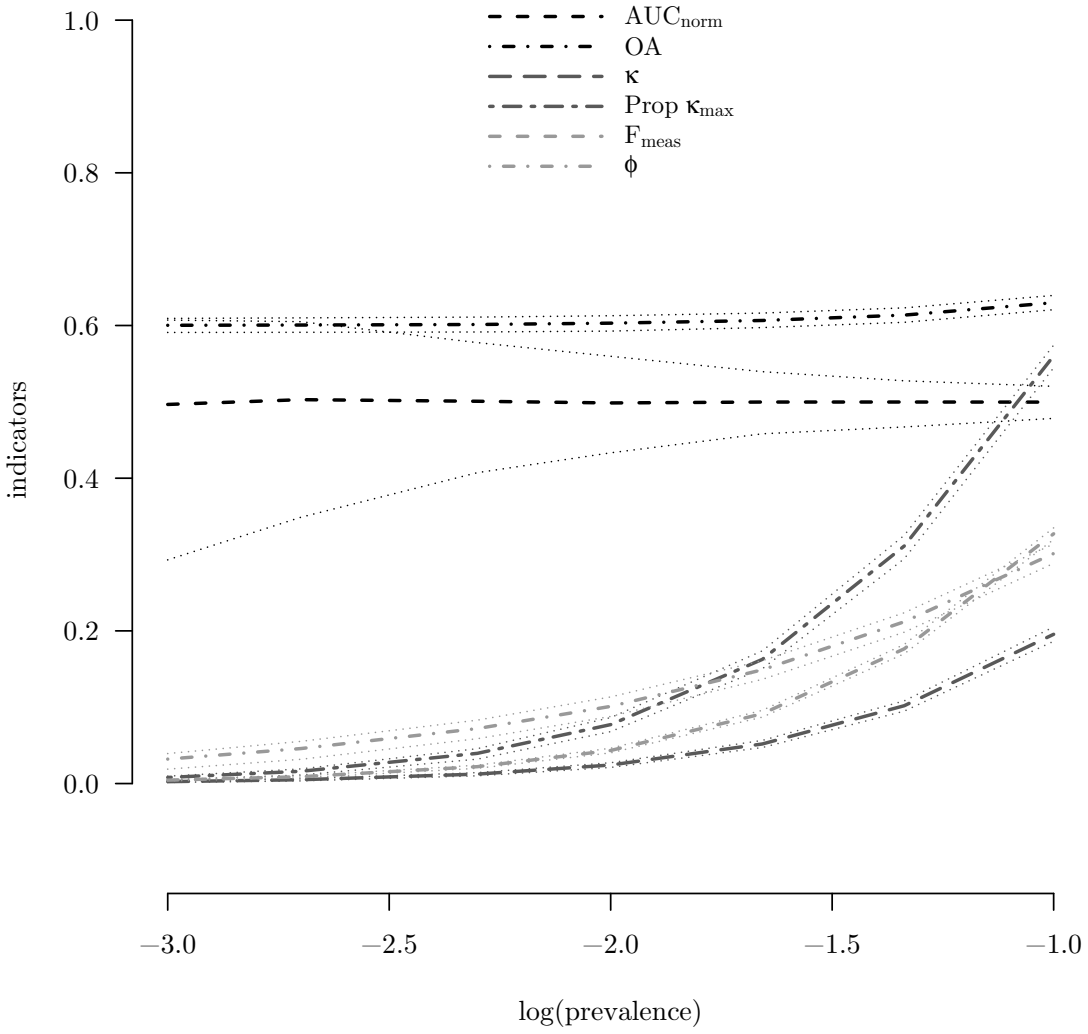


Figure 2.4 – Mean values of indicators *versus* log(prevalence) for sensitivity = 0.9 and specificity = 0.6. AUC_{norm}: normalized area under curve, κ: Cohen's kappa, F_{meas}: F-measure, OA: observed agreement, ϕ: phi coefficient, *Prop* κ_{max}: proportion of maximum kappa (thin dotted lines represent variation intervals)

2.4.1 Main results

This confrontation of indicators brings to light an almost symmetrical behavior of κ with regard to prevalence, with maximum values when frequencies of positive

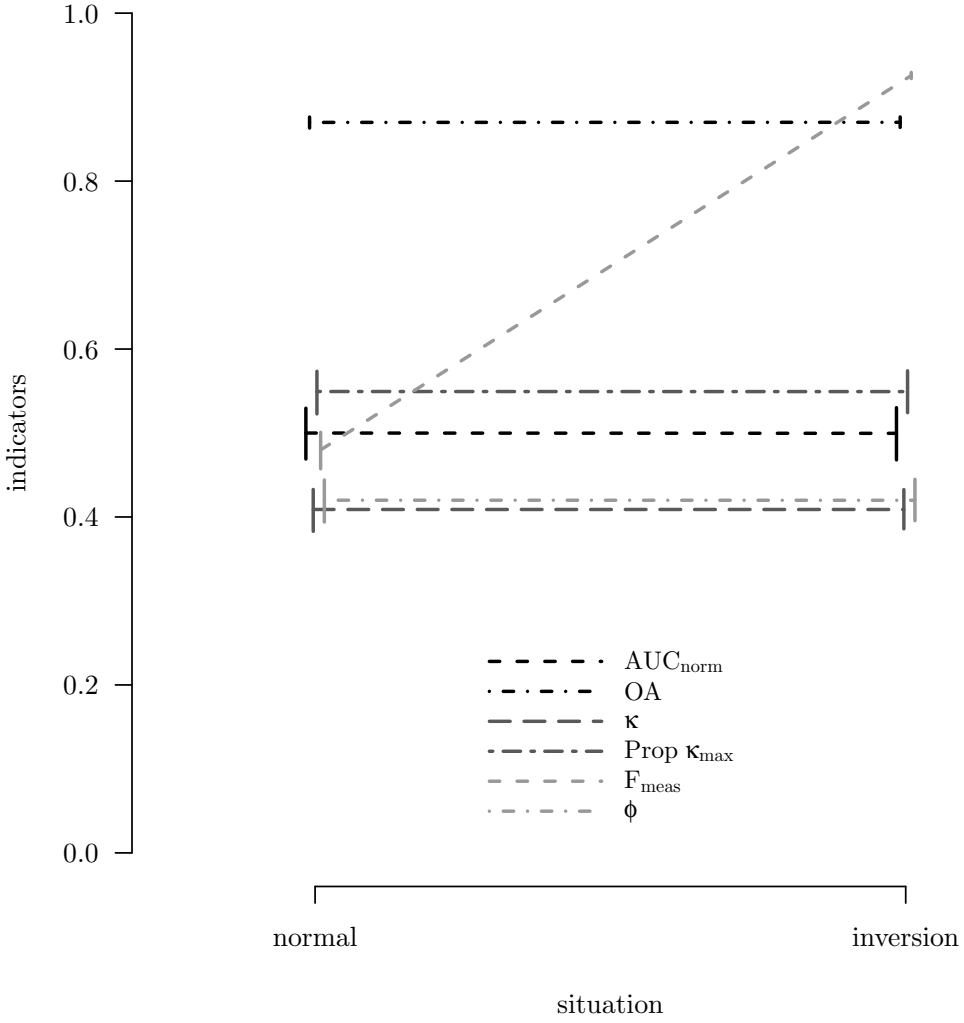


Figure 2.5 – Value of indicators for two opposite situations, "normal" (prevalence = 0.1, sensitivity = 0.6 and specificity = 0.9) and "inversion" (prevalence = 0.9, sensitivity = 0.9 and specificity = 0.6). Mean values are joined by dotted line, variation intervals are represented by vertical lines. AUC_{norm}: normalized area under curve, κ: Cohen's kappa, F_{meas}: F-measure, OA: observed agreement, ϕ: phi coefficient, *Prop* κ_{max}: proportion of maximum kappa

and negative cases are balanced and minimum values for high or low prevalence.

The similar behavior of ϕ can be explained by the fact that ϕ is also a chance-

corrected association coefficient for binary variables [21].

On the contrary, F_{meas} tends to increase steadily and go widely over AUC_{norm} for large prevalence. Thus, F_{meas} is likely to overestimate the diagnostic test performance when prevalence increases because it does not account for true negative cases and it is biased. This finding was highlighted by Powers who pointed out difficulties in understanding the meaning of F_{meas}, and proposed κ as an alternative [20]. Besides, another advantage of κ is its simple application to classification problems with more than two classes [30].

OA is biased and leads to misinterpretation because it is a majority-class classifier [30] and does not account for agreement by chance.

Regarding *Prop* κ_{max}, although the shape of its variations is similar to the one of κ or ϕ, with a maximum for balanced samples and a minimum for extreme prevalence, it seems to be too optimistic, with a maximum that goes over *OA* when *Se* and *Sp* are higher than 0.6 (see appendix B figure B.1), whereas the performance of the test is clearly not so high.

In this experiment, mean values of AUC_{norm} are obviously constant because it directly depends on predefined values of *Se* and *Sp*.

Regarding variation interval, our results show a large interval of AUC_{norm} for extreme prevalences whereas it seems to be constant and thin along prevalence for other indicators. This points out an issue in the interpretation of AUC_{norm} for high or low prevalence. Conversely, AUC_{norm} appears to be close to κ and ϕ in case of balanced samples with a fair variation interval. These observations can obviously be applied to unnormalized AUC.

2.4.2 Interpretation of the case study

Table 2.3 presents values of indicators for the case study.

Table 2.3 – Indicators computed from evaluation of a computerized ECG interpretation reported in Ho *et al.* [22]

rhythm	P^{+++}	$Se^{\S\S}$	$Sp^{\S\S\S}$	AUC_{norm}^{*}	OA^{+}	κ^{**}	$Prop\ \kappa_{max}{}^{\S}$	$F_{meas}{}^{***}$	ϕ^{++}
sinus	0.71	0.355	0.967	0.321	0.533	0.221	0.949	0.518	0.333
atrial fibrillation	0.1	0.927	0.879	0.806	0.884	0.562	0.73	0.622	0.608
atrial flutter	0.04	0.622	0.934	0.553	0.919	0.331	0.394	0.366	0.367

*normalized area under curve, **Cohen's kappa, ***F-measure, +observed agreement, ++phi coefficient, +++prevalence, §proportion of maximum kappa, §§sensitivity, §§§specificity

Results for AFL are represented in figure 2.3 by a vertical line corresponding to the prevalence of this arrhythmia. Thus, for low prevalence, the four indicators κ, *Prop* κ_{max}, ϕ, and F_{meas} are close. The aggregated indicator reported in the paper was OA, however it largely overestimates the diagnostic performance of this arrhythmia.

Available data concerns only one sample and does not enable to observe a variation interval. Nevertheless, in accordance with the simulation results, the real value of AUC_{norm} for AFL diagnosis can stay in the interval from 0.45 to 0.54, approximately.

Regarding diagnosis of AF, κ, ϕ and F_{meas} still agree. Nevertheless, regarding the diagnosis of sinus rhythm, some discrepancies can be observed between F_{meas} and κ and between F_{meas} and ϕ, in spite of a low Se. This is due to the high prevalence of this electrocardiographic status.

According to the above-mentioned conditions of applicability of indicators, κ and ϕ can be considered for the three situations whereas F_{meas} should only be used for AF and AFL detection. Thus, performance of computerized ECG interpretation appears to be higher for AF detection and lower for sinus rhythm diagnosis. However, from an epidemiological perspective, low Se of sinus rhythm detection may not be a serious issue because it only leads to a large amount of false negative (i.e., false positive for abnormal rhythm detection) which can be checked by care providers.

2.4.3 Recommendations

Our results lead to recommendations to help in the choice of a diagnostic test performance indicator to reflect the usability of the test in clinical practice. Different scenarios can be considered:

(i) When dealing with a low prevalence (lower than 0.1), which is a common case in epidemiology, the report of AUC alone should be avoided because of its large variation interval. Others indicators, such as κ, ϕ and F_{meas} should be preferred. *Prop* κ_{max} should be used carefully because its value is systematically above κ and ϕ when the prevalence becomes higher than about 0.05. Moreover, this indicator seems to provide misleading results for prevalence around 0.5.

(ii) When the prevalence is high, one should wonder if the health status to be diagnosed and its opposite have not been inverted. In this case, F_{meas} is not reliable and should not been used, whereas κ and ϕ are strictly insensitive to this situation.

(iii) The presence of a gold standard should not lay aside the κ coefficient. The

use of κ may indeed be questioned in this context because this indicator is usually used in other circumstances and has been designed for evaluation of agreement among two or more judges, without stating who is the reference, contrary to the presence of a gold standard. However, the values of κ are close to ϕ and also to F_{meas} for low prevalence, so it seems to reflect well the correlation which may exists between data. From our point of view, it is also fair to use κ as an indicator of diagnostic test performance with comparison to a gold standard test. In fact, it is used in this way in several studies (example in *Chiu et al.* [31]).

Report of study results

The results reported in a study should reflect the reproducibility of the test in clinical practice. Besides important methodological considerations, especially the characteristics of the studied population and the choice of the gold standard, the results have to deal with two concepts [7, 32]:

diagnostic accuracy, which is an intrinsic feature of the test, reflecting the measurement error, accounting for intra-subject variability,

reliability, which accounts for inter-subject variability to be able to discriminate subjects.

Thus, Se and Sp as well as AUC should be reported as markers of accuracy. Moreover, ROC curve, beyond the computation of AUC, is useful in defining a cut-off value to transform a quantitative variable into a binary variable. However, the large variation interval of AUC should be kept in mind and should avoid making strong assumptions about the test efficiency.

Nevertheless, these indicators are inadequate to reflect the reliability of the

test to discriminate the diagnostic classes especially for rare diseases. In this case, κ tends to be used above all because it is likely to be better understood than ϕ which is not well known among physicians, and can be applied more widely than F_{meas}.

Finally, indicators may not be sufficient to reflect the entire information about the performance of a diagnostic test, and raw data of contingency table should be displayed with the results [7].

2.5 Conclusions

This simulation-based study gives an overview of the behavior of most common aggregated indicators of performance of a diagnostic test, and provides some clues to help the researcher in choosing the most suitable indicator in the context of his study.

Chapitre 3

Discussion

3.1 Principaux résultats

Ce travail a permis de confronter les indicateurs synthétiques de performance d'un test diagnostique dans différentes situations.

3.1.1 Concernant l'effet du taux de prévalence

La valeur moyenne de l'AUC_{norm} est indépendante du taux de prévalence dans la mesure où elle ne dépend que du couple $\{Se, Sp\}$ fixé par l'expérience, et reflète donc la performance intrinsèque du test.

On observe un comportement quasi symétrique du κ avec, pour des caractéristiques intrinsèques données, des valeurs optimales quand les effectifs de chaque classe (malades et non malades) sont équilibrés, c'est-à-dire pour des taux de prévalence d'environ 0,5, et des valeurs minimales quand une classe prédomine l'autre, c'est-à-dire pour des taux de prévalence élevés ou bas. On voit également que la valeur maximale de cet indicateur correspond à peu près à l'AUC_{norm}. Le com-

portement du coefficient ϕ s'avère très proche, ce qui peut être expliqué par le fait que, à l'instar du κ, ϕ est aussi un coefficient d'association ajustée sur le hasard pour les variables binaires [21].

Au contraire, le comportement de la F_{mes} se montre très asymétrique, avec des valeurs constamment maximales lorsque le taux de prévalence est élevé. On observe d'ailleurs que la F_{mes} dépasse l'AUC_{norm} dès que le taux de prévalence n'est plus faible, c'est-à-dire supérieur à 0,05 à 0,2 selon la performance intrinsèque du test. On peut donc avancer que cet indicateur a tendance à surestimer la performance du test si le taux de prévalence n'est pas bas. Ceci est du au fait que la F_{mes} ne prend pas en compte les vrais négatifs. Ce comportement a été décrit par Powers qui propose le κ comme une alternative à la F_{mes} [20].

La CO se situe invariablement au dessus de l'AUC_{norm} et semble donc surestimer la performance du test, ce qui s'explique par l'absence de prise en compte de la part de concordance due au hasard. S'appuyant sur la notation de la figure 1.2, le développement suivant montre que la CO est toujours supérieure ou égale au κ.

Démonstration. Sachant que toutes les quantités sont positives :

$$
\begin{aligned}
& D \geq C \\
\Leftrightarrow\ & BD \geq BC \\
\Leftrightarrow\ & BD + CD \geq BC + CD \\
\Leftrightarrow\ & \frac{B+C}{B+D} \geq \frac{C}{D} \\
\text{donc}\ & CO \geq \kappa
\end{aligned}
$$

□

Enfin, concernant la *Prop* κ_{max}, si la forme de son évolution avec le taux de prévalence ressemble à celle du κ, ses valeurs sont en général beaucoup plus optimistes dès que le taux de prévalence s'éloigne des bornes inférieures et supérieures, jusqu'à dépasser la F_{mes} et la CO. Par exemple, pour le cas où Se et Sp sont de 0.8, la valeur du *Prop* κ_{max} correspondant à un taux de prévalence de 0,5 est de 0,95, ce qui surestime la performance du test d'une façon évidente (voir annexe B figure B.1).

3.1.2 Lorsque le taux de prévalence est faible ($\leq$ 0,1)

Les indicateurs κ, *Prop* κ_{max}, ϕ et F_{mes} sont assez proches, avec un κ constamment moins optimiste que les autres. Là encore, la CO dépasse largement l'AUC_{norm}.

3.1.3 Dans la situation où les diagnostics « positif » et « négatif » sont inversés

La F_{mes} est le seul indicateur sensible à cette inversion de situation. Ceci se comprend aisément par l'étude de son expression mathématique (équation 1.12 page 10) : alors que les autres indicateurs prennent en compte de façon symétrique les cas « positifs » (VP, FP) et « négatifs » (VN, FN), il n'en est pas de même pour la F_{mes} où la quantité VN n'apparaît pas. L'inversion de situation entraîne une inversion des valeurs VP et VN d'où une modification de la F_{mes}.

3.1.4 Concernant l'intervalle de fluctuation

Seule l'AUC_{norm} montre des variations notables en fonction du taux de prévalence, avec un élargissement important de cet intervalle dès que le taux de

prévalence est inférieur à 0,01 voire 0,05. L'intervalle de fluctuation des autres indicateurs reste fin et constant avec le taux de prévalence.

3.2 Forces et limites

L'originalité de ce travail tient à la comparaison de plusieurs indicateurs synthétiques par une méthode exclusivement graphique permettant une visualisation facile de leur comportement. Si une étude analytique des expressions mathématiques des indicateurs pourrait apporter des explications sur les comportements observés, elle ne serait pas réalisable pour tous les indicateurs simultanément, et ne permettrait pas une interprétation facilement compréhensible par les chercheurs et cliniciens.

La représentation des intervalles de fluctuation pour une taille d'échantillon arbitrairement choisie est un autre intérêt de ce travail. Elle permet aussi de rappeler au lecteur qu'une valeur d'un indicateur donnée dans une étude est une estimation ponctuelle entachée d'erreur, comme tout résultat statistique. Ce point est particulièrement critique pour l'AUC, comme nous l'avons illustré.

Le fait que la taille des échantillons soit fixe pour toutes les simulations (10^4) est une limite de notre approche. En effet, la largeur des intervalles de variation dépend de cette taille. Néanmoins, la comparaison des intervalles de fluctuation entre les différents indicateurs reste valable dans la mesure où ils ont tous été déterminés dans les mêmes conditions.

3.3 Interprétation du cas d'étude

Les indicateurs étudiés dans ce travail ont été calculés à partir des données brutes rapportées dans l'étude de Ho *et al.* [22] (tableau 3.1). La valeur des indicateurs pour le diagnostic du flutter auriculaire (FLA) sont représentées également sur la figure 2.3 page 29.

TABLEAU 3.1 – Indicateurs de performance d'un système d'interprétation automatisée d'électrocardiogramme pour trois rythmes cardiaques [22]. FA : fibrillation auriculaire, FLA : flutter auriculaire, AUC_{norm} : aire sous la courbe normalisée, κ : coefficient kappa de Cohen, F_{mes} : F-mesure, CO : concordance observée, ϕ : coefficient phi, P : taux de prévalence, *Prop* κ_{max} : proportion du kappa maximal, Se : sensibilité, Sp : spécificité

Rythme	P	Se	Sp	AUC_{norm}	CO	κ	*Prop* κ_{max}	F_{mes}	ϕ
sinusal	0,71	0,355	0,967	0,321	0,533	0,221	0,949	0,518	0,333
FA	0,1	0,927	0,879	0,806	0,884	0,562	0,73	0,622	0,608
FLA	0,04	0,622	0,93	0,553	0,919	0,331	0,394	0,366	0,367

Concernant le diagnostic des deux arythmies, à savoir la fibrillation auriculaire (FA) et le FLA, on observe que les quatre indicateurs κ, *Prop* κ_{max}, ϕ et F_{mes} sont concordants pour accorder une meilleure performance pour le diagnostic de la FA. La *Prop* κ_{max} se montre plus optimiste que les trois autres, et cela est concordant avec le résultats des simulations. Un faible taux de prévalence de ces arythmies (de moins de 0,1) explique que les valeurs de la F_{mes} ne soient pas plus élevées. L'AUC_{norm} va dans le même sens, même si elle donne l'impression de meilleures performances, puisqu'elle ne prend pas en compte le contexte d'application du test qui dans ce cas n'est pas « optimal » (un contexte « optimal » serait un taux de prévalence proche de 0,5). L'indicateur synthétique renseigné par les auteur de cette étude nous servant d'exemple est la CO. Or, cet indicateur aboutit à des conclusions inverses, avec d'une part de bonnes performances pour le diagnostic

des deux arythmies, et d'autre part une performance meilleure pour le FLA. Ce résultat est du à un taux de prévalence faible de la pathologie, masquant le déficit en sensibilité, et un taux de prévalence de cas sans cette pathologie élevé, biaisant le résultat en faveur de la spécificité.

Ces données ne concernant qu'un échantillons de cas, on ne dispose pas de l'intervalle de variation de l'AUC_{norm}. Cependant, en se référant à la figure 2.3 page 29, on voit que, pour ce taux de prévalence (0,04), cet intervalle est plus important que pour les autres indicateurs. Nous ne pouvons déterminer ses limites d'après les simulations, puisque le nombre d'individus de l'étude est d'environ 200 000 individus alors qu'il n'est que de 10 000 dans les simulations.

Concernant le diagnostic du statut « rythme sinusal », on remarque une valeur de F_{mes} plus élevée que les indicateurs κ et ϕ. Cette observation est concordante avec les résultats des simulations : le rythme sinusal étant un diagnostic fréquent (taux de prévalence = 0,7), la F_{mes} surestime la performance du test. On peut considérer qu'il s'agit là d'une inversion de diagnostic : la pathologie est le fait de ne pas avoir de rythme sinusal. Pourtant, la présence d'un rythme sinusal étant une assertion classique dans l'analyse d'un ECG, c'est le message « rythme sinusal » (fréquent et physiologique) qui est classiquement évalué, et non « rythme non sinusal » (plus rare et pathologique). On observe également que la *Prop* κ_{max} paraît ici inadaptée, avec une valeur excellente alors que les autres indicateurs vont dans le sens d'une mauvaise performance globale.

Cette étude de cas met donc bien en évidence qu'une utilisation du κ et du ϕ est possible dans les trois situations, alors que la F_{mes} et la *Prop* κ_{max} ne trouvent leur place que quand le taux de prévalence est faible (ou élevée pour la *Prop* κ_{max}, d'après les simulations) et que la *CO* peut aboutir à des conclusions erronées.

3.4 Recommandations

Les résultats de notre travail amènent à proposer des recommandations pour aider le chercheur à choisir l'indicateur synthétique qui reflétera au mieux la performance d'un indicateur diagnostique dans son contexte, donc son utilisabilité. Précisons que nous étendons le comportement observé de l'AUC_{norm} à l'AUC.

3.4.1 Différents scénarios

1. Lorsque le taux de prévalence de la pathologie à diagnostiquer est faible (inférieure ou égale à 0,1), ce qui est très courant, l'utilisation de l'AUC seule est à éviter du fait d'un intervalle de fluctuation potentiellement important. De plus, rappelons que l'AUC ne prend pas en compte le taux de prévalence [4, 14]. Les indicateurs κ, ϕ et F_{mes} sont probablement à préférer. La *Prop* κ_{max} pourrait avoir une place, bien que pour des taux de prévalence proches de 0,1, ses valeurs apparaissent toujours supérieures aux κ et ϕ surtout si la *Sp* est basse (voir les résultats graphiques pour les différents couples $\{Se, Sp\}$ figure B.2 page 65).
2. Lorsque le taux de prévalence est élevé, il est important de se demander si les statuts « pathologie » (état à détecter) et « sain » n'ont pas été inversés. Dans ce cas, l'utilisation de la F_{mes} ne paraît pas souhaitable dans la mesure où cet indicateur semble trop optimiste.
3. La présence d'un *gold standard*, ce qui est la règle dans les études d'évaluation d'un indicateur diagnostique, pourrait faire penser que l'utilisation du κ n'est pas convenable. En effet, le κ est un indicateur qui a été développé pour évaluer la concordance entre deux juges, sans spécifier que l'un constitue

une référence. Néanmoins, nos résultats montrent que ses valeurs sont très proches du coefficient ϕ, et de la F_{mes} quand le taux de prévalence est faible. D'autre part, une des hypothèses sous-tendant le κ est l'indépendance des distributions des deux évaluations, ce qui correspond à la réalité d'une étude où la méthodologie exige que les deux évaluations, le *gold standard* et le test à l'épreuve, soient menés indépendamment. À l'inverse, la F_{mes} suppose que les deux évaluations proviennent d'une même distribution [20]. Ainsi, nous pensons qu'il est licite d'utiliser le κ dans ce contexte, et ceci est appliqué dans certaines études (exemple avec Chiu *et al.* pour l'évaluation d'un système d'interprétation automatisée d'ECG en pédiatre [31]) mais aussi dans le domaine du *machine learning* pour comparer différents tests [20].

3.4.2 Communication des résultats d'une étude diagnostique

Les résultats rapportés dans la publication d'une étude doivent permettre d'évaluer l'applicabilité du test en pratique clinique.

Cette applicabilité dépend bien sûr de caractéristiques méthodologiques, à savoir les conditions dans laquelle l'étude est réalisée, pouvant être source de variations dans les résultats du test, mais aussi la présence de biais [33]. Les principales conditions de réalisation à confronter avec la réalité clinique sont : la population sur laquelle le test a été validé (censée être représentative de la population cible du test), le test et ses conditions d'application, le *gold standard* choisi et les critères de jugements [34]. L'outil QUADAS-2 (*Quality Assessment of Diagnostic Accuracy Studies*) est un des outils destinés aux revues systématiques d'études de performance d'un test diagnostique, et propose une démarche structurée, chaque

domaine d'analyse étant évalué en termes de risque de biais et d'applicabilité [26].

Concernant les indicateurs de performance, la reproductibilité implique deux concepts [7, 32] :

la précision, caractéristique intrinsèque au test, qui reflète l'erreur de mesure et la variabilité intra-sujet,

la fiabilité, qui a trait à la variabilité inter-sujets et la capacité de discrimination des sujets.

Ainsi, la Se, la Sp et l'AUC doivent être renseignées en tant que marqueurs de précision. Par ailleurs, la construction d'une courbe ROC, d'où découle l'AUC, permet aussi de choisir une valeur seuil dans le cas où l'on cherche à transformer un résultat quantitatif du test en un résultat binaire (diagnostic confirmé ou non). Néanmoins, il faut se souvenir d'un intervalle de fluctuation de l'AUC potentiellement large, qu'il faut rapporter [5]. Czodrowski précise aussi qu'il ne faut pas se limiter à rapporter la Se, la VPP et la CO qui sont biaisées vers la classe prédominante [30]. En effet, baser la prédiction uniquement sur la modalité la plus fréquente du test suffit à obtenir des bons résultats.

Néanmoins, ces indicateurs ne sont pas à même de donner une indication sur la fiabilité du test en pratique, permettant son utilisabilité [5]. Nous suggérons pour cela l'utilisation du κ. En effet, plus largement utilisable que la F_{mes} (notamment pour les taux de prévalence élevés), largement moins optimiste que la CO, et plus connu que le ϕ parmi les cliniciens, sa signification est aussi plus compréhensible et démonstrative (*via* la figure 1.2 page 9 par exemple).

Quoiqu'il en soit, l'entièreté de l'information sur la performance d'un test diagnostique, contenu dans le tableau de contingence, ne peut être résumée en un seul

indicateur. Distinguer les performances spécifiques aux cas positifs et négatifs, *via* les couples d'indicateurs partiels présentés en introduction (section 1.2.1), garde tout son sens puisqu'il convient de prendre en compte les conséquences des faux positifs et faux négatifs [5]. Ainsi, la mauvaise *Se* de la détection du rythme sinusal dans notre cas d'étude (tableau 3.1 page 41) peut n'avoir que peu d'impact négatif en pratique dans la mesure où les ECG classés « pathologiques » à tort seront probablement revus par des cliniciens avant l'instauration d'une thérapeutique. La caractéristique principale demandée au système est une bonne *Se* non pas pour le rythme sinusal, qui n'est pas la pathologie mais la norme, mais pour les arythmies afin d'éviter un défaut de prise en charge des patients atteints. Or les indicateurs que nous avons étudiés pondèrent de la même façon les erreurs par excès et celles par omission [11]. Ainsi, des auteurs recommandent de ne pas se limiter à fournir la valeur des indicateurs mais aussi de présenter le tableau de contingence duquel sont dérivés les calculs [7].

Conclusion

Ce travail a permis de décrire le comportement d'indicateurs synthétiques de performance d'un test diagnostique et de fournir au chercheur des recommandations quant à leur utilisation afin de choisir l'indicateur le plus adapté à la situation de l'étude. L'expression des résultats sous forme de comparaisons graphiques permet de se rendre compte simplement des points essentiels pour la compréhension de ces indicateurs. Une focalisation sur les situations de taux de prévalence faible correspond à l'application réelle de nombreux tests diagnostiques en pratique clinique et permet de montrer qu'il ne faut pas se satisfaire seulement des indicateurs de précision habituels que sont la sensibilité, la spécificité et l'AUC qui en est dérivée. Si la situation d'inversion des états positifs et négatifs est plus rarement rencontrée, cette étude démontre qu'elle peut avoir des conséquences inattendues si la F_{mes} était l'indicateur choisi *a priori*.

Dans l'objectif de refléter l'utilisabilité d'un test diagnostique, nos résultats nous amènent à privilégier le coefficient kappa de Cohen qui semble être l'indicateur le plus largement applicable et le moins à risque d'excès d'optimisme. Nous avons également montré que sa signification est intuitive et peut être expliquée graphiquement.

Il faut néanmoins se souvenir que la valeur d'un indicateur, fût-ce le κ, ne

peut fournir toute l'information disponible sur la performance du test disponible à partir du tableau de contingence, et que l'étude des effectifs présents dans ce dernier garde son importance.

Bibliographie

[1] Habbema JD. Clinical decision theory : the threshold concept. Neth J Med. 1995 ;47(6) :302–307.

[2] Versi E. "Gold standard" is an appropriate term. BMJ. 1992 ;305(6846) :187. Disponible sur : http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1883235/.

[3] Duggan PF. Time to abolish “gold standard”. BMJ. 1992 ;304(6841) :1568–1569. Disponible sur : http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1882438/.

[4] Lasko TA, Bhagwat JG, Zou KH, Ohno-Machado L. The use of receiver operating characteristic curves in biomedical informatics. J Biomed Inform. 2005 ;38(5) :404–415.

[5] Bossuyt P, Davenport C, Deeks J, Hyde C, Leeflang M, Scholten R. Handbook for Diagnostic Test Accuracy Reviews | Diagnostic Test Accuracy Working Group. The Cochrane Collaboration ; 2013. Disponible sur : http://srdta.cochrane.org/handbook-dta-reviews.

[6] Altman DG, Bland JM. Diagnostic tests 1 : sensitivity and specificity. BMJ. 1994 ;308(6943) :1552. Disponible sur : http://www.ncbi.nlm.nih.gov/pmc/

articles/PMC2540489/.

[7] Hernaez R. Reliability and agreement studies : a guide for clinical investigators. Gut. 2015 ;64(7) :1018–1027.

[8] Akobeng AK. Understanding diagnostic tests 1 : sensitivity, specificity and predictive values. Acta Pædiatrica. 2007 ;96(3) :338–341. Disponible sur : http://onlinelibrary.wiley.com/doi/10.1111/j.1651-2227.2006.00180.x/abstract.

[9] Altman DG, Bland JM. Diagnostic tests 2 : predictive values. BMJ. 1994 ;309(6947) :102. Disponible sur : http://www.bmj.com/content/309/6947/102.1.

[10] Mandrekar JN. Simple statistical measures for diagnostic accuracy assessment. J Thorac Oncol. 2010 ;5(6) :763–764.

[11] Lobo JM, Jiménez-Valverde A, Real R. AUC : a misleading measure of the performance of predictive distribution models. Global Ecology and Biogeography. 2008 ;17(2) :145–151. Disponible sur : http://onlinelibrary.wiley.com/doi/10.1111/j.1466-8238.2007.00358.x/abstract.

[12] Deeks JJ, Altman DG. Diagnostic tests 4 : likelihood ratios. BMJ. 2004 ;329(7458) :168–169. Disponible sur : http://www.bmj.com/content/329/7458/168.

[13] Altman DG, Bland JM. Diagnostic tests 3 : receiver operating characteristic plots. BMJ. 1994 ;309(6948) :188. Disponible sur : http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2540706/.

[14] Centor RM. Signal Detectability The Use of ROC Curves and Their Analyses. Med Decis Making. 1991 ;11(2) :102–106. Disponible sur : http://mdm.sagepub.com/content/11/2/102.

[15] Powers DM. Evaluation : from precision, recall and F-measure to ROC, informedness, markedness and correlation. Journal of Machine Learning Techn ologies. 2011 ;2(1) :37–63. Disponible sur : http://dspace.flinders.edu.au/xmlui/handle/2328/27165.

[16] Fawcett T. An introduction to ROC analysis. Pattern recognition letters. 2006 ;27(8) :861–874. Disponible sur : http://www.sciencedirect.com/science/article/pii/S016786550500303X.

[17] Cohen J. A Coefficient of Agreement for Nominal Scales. Educational and Psychological Measurement. 1960 ;20(1) :37–46. Disponible sur : http://epm.sagepub.com/content/20/1/37.

[18] Sim J, Wright CC. The Kappa Statistic in Reliability Studies : Use, Interpretation, and Sample Size Requirements. PHYS THER. 2005 ;85(3) :257–268. Disponible sur : http://ptjournal.apta.org/content/85/3/257.

[19] Hoehler FK. Bias and prevalence effects on kappa viewed in terms of sensitivity and specificity. J Clin Epidemiol. 2000 ;53(5) :499–503.

[20] Powers D. What the F-measure doesn't measure. Australia : Flinders University ; 2014. KIT-14-001. Disponible sur : https://www.academia.edu/11523406/What_the_F-measure_doesnt_measure.

[21] Warrens MJ. Chance-corrected measures for 2 x 2 tables that coincide with weighted kappa. Br J Math Stat Psychol. 2011 ;64(Pt 2) :355–365.

[22] Ho TW, Huang CW, Lin CM, Lai F, Ding JJ, Ho YL, *et al.* A telesurveillance system with automatic electrocardiogram interpretation based on support vector machine and rule-based processing. JMIR Med Inform. 2015 ;3(2) :e21.

[23] Umesh UN, Peterson RA, Sauber MH. Interjudge agreement and the maximum value of kappa. Educational and Psychological Measurement. 1989 ;49(4) :835–850. Disponible sur : `http://epm.sagepub.com/content/49/4/835`.

[24] Sasaki Y. The truth of the F-measure ; 2007. Disponible sur : `http://www.cs.odu.edu/~mukka/cs795sum09dm/Lecturenotes/Day3/F-measure-YS-26Oct07.pdf`.

[25] Hripcsak G, Rothschild AS. Agreement, the F-Measure, and Reliability in Information Retrieval. J Am Med Inform Assoc. 2005 ;12(3) :296–298. Disponible sur : `http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1090460/`.

[26] Whiting PF, Rutjes AWS, Westwood ME, Mallett S, Deeks JJ, Reitsma JB, *et al.* QUADAS-2 : a revised tool for the quality assessment of diagnostic accuracy studies. Ann Intern Med. 2011 ;155(8) :529–536.

[27] Hripcsak G, Heitjan DF. Measuring agreement in medical informatics reliability studies. J Biomed Inform. 2002 ;35(2) :99–110.

[28] Warrens MJ. On association coefficients for 2 x 2 tables and properties that do not depend on the marginal distributions. Psychometrika. 2008 ;73(4) :777–789. Disponible sur : `http://link.springer.com/article/10.1007/s11336-008-9070-3`.

[29] R Core Team. R : A Language and Environment for Statistical Computing. Vienna, Austria ; 2015. Disponible sur : `https://www.R-project.org/`.

[30] Czodrowski P. Count on kappa. J Comput Aided Mol Des. 2014 ;28(11) :1049–1055.

[31] Chiu CC, Hamilton RM, Gow RM, Kirsh JA, McCrindle BW. Evaluation of computerized interpretation of the pediatric electrocardiogram. J Electrocardiol. 2007 ;40(2) :139–143.

[32] de Vet HCW, Terwee CB, Knol DL, Bouter LM. When to use agreement versus reliability measures. J Clin Epidemiol. 2006 ;59(10) :1033–1039.

[33] Whiting PF, Rutjes AWS, Westwood ME, Mallett S, QUADAS-2 Steering Group. A systematic review classifies sources of bias and variation in diagnostic test accuracy studies. J Clin Epidemiol. 2013 ;66(10) :1093–1104.

[34] Schmidt RL, Factor RE. Understanding sources of bias in diagnostic accuracy studies. Arch Pathol Lab Med. 2013 ;137(4) :558–565.

Annexe A

Script

Cette section contient le script en langage R [29] utilisé pour réaliser les simulations informatiques et les sorties graphiques.

```
##################
## Parameters ##
##################


## Set of parameters

## Prevalence (P)
list_P <- c(round(c(10^-3, 10^-(2+2/3), 10^-(2+1/3), 10^-2,
10^-(1+2/3), 10^-(1+1/3), 10^-1), 3), 0.2, 0.3, 0.4, 0.5, 0.6, 0.7,
0.8, 0.9, 0.99)

## Sensitivity (Se)
list_Se <- seq(0.5, 0.9, 0.1)

## Specificity (Sp)
list_Sp <- seq(0.5, 0.9, 0.1)


## Number of simulations for each combination of P, Se, Sp
nb_simulations <- 1000

## Population size
N <- 10000


###################
## Simulations ##
###################
```

```
31
## Simulations for each combination of P, Se and Sp (each saved in a
## CSV file named "simulations_P_Se_Sp.csv")
for (P in list_P) {
    for (Se in list_Se) {
36      for (Sp in list_Sp) {

            ## CSV file to save the simulations for this combination
            ## of parameters
            result_file <- paste("simulations_", P, "_", Se, "_", Sp,
41                               ".csv", sep = "")


            ## Headers of columns: true positive, false positive,
            ## false negative, true negative
46          cat("TP,FP,FN,TN", "\n", sep = "", file = result_file)

            for (i in 1:nb_simulations) {
                ## Index of subjects
                index <- 1:N
51

                ## Real values
                ## Index of real positive cases
                index_event <- sample(index, P*N, replace = F)
56              ## Index of real negative cases
                index_no_event <-
                    index[which(! index %in% index_event)]

                ## Observed (real) values in the population, with a
61              ## prevalence of event = P
                x_obs <- factor(rep(0,N), levels=c(0,1))
                x_obs[index_event] <- 1


66              ## Predicted values
                x_pred <- factor(rep(0,N), levels=c(0,1))

                ## For real positive cases: probability of positive
                ## predicted value = Se
71              x_pred[index_event] <-
                    as.integer(runif(length(index_event),0,1) < Se)
                ## For real negative cases: probability of positive
                ## predicted value = 1 - Sp
                x_pred[index_no_event] <-
76                  as.integer(runif(length(index_no_event),0,1) > Sp)


                ## Construction of the contingency matrix
```

```
                contingency_table <- as.matrix(table(x_pred,x_obs))

                ## save in the CSV file, to prevent RAM overload
                cat(paste(contingency_table[2,2], # TP
                          contingency_table[2,1], # FP
                          contingency_table[1,2], # FN
                          contingency_table[1,1], # TN
                          sep=","),
                    "\n", sep = "", file = result_file, append=T)
            }
        }
    }
}



#####################################
## Computation of indicators ##
#####################################

## Dataframe to gather all computed indicators
indicators <- data.frame(P = numeric(0), # prevalence
                         Se = numeric(0), # observed sensitivity
                         Sp = numeric(0), # observed specificity
                         AUCnorm = numeric(0), # area under curve
                         Kappa = numeric(0), # Cohen's kappa
                         OA = numeric(0),    # observed agreement
                         PropKappaMax = numeric(0), # proportion of maximum kappa
                         Fmes = numeric(0),         # F-measure
                         phi = numeric(0)          # phi coefficient
                         )

## Computation for each combination of P, Se and Sp
for (P in list_P) {
    for (Se in list_Se) {
        for (Sp in list_Sp) {

            simulation_file <- paste("simulations_", P, "_", Se, "_",
                                     Sp, ".csv", sep = "")

            data <- read.csv(simulation_file, header=T)

            # Observed sensitivity
            data$SeMes <- data$TP / (data$TP + data$FN)

            # Observed specificity
            data$SpMes <- data$TN / (data$TN + data$FP)

            ## Normalized area under curve
```

```
            data$AUCnorm <- (data$SeMes * data$SpMes + data$SeMes *
                              (1 - data$SpMes) / 2 +
                              (1 - data$SeMes) * data$SpMes / 2) * 2 - 1

            ## Cohen's kappa
            data$Kappa <- (2 * (data$TP * data$TN - data$FP * data$FN)) /
                ((data$TP + data$FN) * (data$FN + data$TN) + (
                    data$FP + data$TN) * (data$TP + data$FP))

            ## Observed agreement
            data$OA <- (data$TP + data$TN) / (data$TP + data$FP +
                                              data$FN + data$TN)

            ## Proportion of maximum kappa
            data$PropKappaMax <- data$Kappa /
                ((data$OA)^2 / ((1 - data$OA)^2 + 1))

            ## F-measure
            data$Fmes <- 2 * data$TP / (2 * data$TP + data$FP +
                                        data$FN)

            ## Phi coefficient
            data$phi <-
                (data$TP/N - (data$TP/N + data$FP/N) *
                 (data$TP/N + data$FN/N))/sqrt((data$TP/N + data$FP/N) * (
                     data$TN/N + data$FN/N) * (data$TP/N + data$FN/N) * (
                         data$FP/N + data$TN/N))

            indicators <- rbind(indicators,
                                data.frame(P = P,
                                           Se = Se,
                                           Sp = Sp,
                                           AUCnorm = data$AUCnorm,
                                           Kappa = data$Kappa,
                                           OA = data$OA,
                                           PropKappaMax = data$PropKappaMax,
                                           Fmes = data$Fmes,
                                           phi = data$phi
                                           )
                                )
        }
    }
}


############################
## Plot of indicators ##
############################
```

```
## Parameters for graphical representation of each indicator
indics <- data.frame(name = c("AUCnorm", "OA", "Kappa",
                              "PropKappaMax", "Fmes", "phi"),
                     color = c("black","black",grey(.35),grey(.35),
                               grey(.6),grey(.6)),
                     line_type = c(2,4,5,6,2,4), # line type
                     stringsAsFactors=F
                     )

## Text for legends of graphics
text.legends <- c(expression(AUC[norm]), "OA", expression(kappa),
                  expression(paste("Prop ", kappa[max])),
                  expression(F[meas]), expression(phi))

## Function to compute mean and variation interval
aggreg.indic.P <- function(df, indic, log_indic= 0, log_P = 0) {
    if (log_indic == 0) {
        moy <- tapply(df[,indic], df["P"], mean)
        quant025 <- tapply(df[,indic], df["P"], quantile, probs=.025)
        quant975 <- tapply(df[,indic], df["P"], quantile, probs=.975)
    } else {
        moy <- tapply(log10(df[,indic]), df["P"], mean)
        quant025 <- tapply(log10(df[,indic]), df["P"], quantile, probs=.025)
        quant975 <- tapply(log10(df[,indic]), df["P"], quantile, probs=.975)
    }

    P <- sort(unique(df["P"]))
    if (log_P != 0) {
        P <- log10(P)
    }

    return (as.data.frame(cbind(P, moy, quant025, quant975)))
}


## For all ranges of prevalences

for (Se in list_Se) {
    for (Sp in list_Sp) {

        indic_SeSp <- indicators[which(indicators$Se == Se
                                       & indicators$Sp == Sp),]

        par (mar=c(5,4,4,6)+.1, xpd=T, las=1, bty="n")

        y_min <- min(c(0, indic_SeSp$AUCnorm, indic_SeSp$Kappa,
                       indic_SeSp$Co, indic_SeSp$Fmes, indic_SeSp$phi))-0.1
```

```
        plot(c(0,1), c(y_min,1),
             xlim=c(0,1),
             ylim=c(y_min,1),
             type="n",
             main= paste("Se = ", Se, ", Sp = ", Sp, sep=""),
             xlab="prevalence",
             ylab="indicators")


        for (indic in  indics$name) {
            color <- indics[which(indics$name == indic), "color"]
            aggreg_indic <- aggreg.indic.P(indic_SeSp, indic, 0, 0)
            lines(aggreg_indic$P, aggreg_indic$moy, type="l",
                  col=color,
                  lty=indics[which(indics$name == indic), "line_type"],
                  lwd=2)
            lines(aggreg_indic$P, aggreg_indic$quant025, type="l",
                  col=color, lty=3)
            lines(aggreg_indic$P, aggreg_indic$quant975, type="l",
                  col=color, lty=3)
        }

        legend("topright", legend=text.legends, col=indics$color,
               lty=indics$line_type, lwd=2, inset=c(-0.2, 0),
               box.lty=0, cex=.8, seg.len = 4)
    }
}


## For low prevalences

for (Se in list_Se) {
    for (Sp in list_Sp) {

        indic_SeSp <-
            indicators[which(indicators$Se == Se & indicators$Sp == Sp
                             & indicators$P <= 0.1),]

        par (mar=c(5,4,4,6)+.1, xpd=T, las=1, bty="n")

        x_min <- min(log10(indic_SeSp$P))
        x_max <- max(log10(indic_SeSp$P))

        y_min <- min(c(0, indic_SeSp$AUCnorm, indic_SeSp$Kappa,
                       indic_SeSp$Co, indic_SeSp$Fmes, indic_SeSp$phi))-0.1

        plot(c(x_min,x_max), c(y_min,1),
             xlim=c(x_min,x_max),
```

```
276             ylim=c(y_min,1),
                type="n",
                main=paste("Se = ", Se, ", Sp = ", Sp, sep=""),
                xlab="log(prevalence)",
                ylab="indicators")
281
        for (indic in  indics$name) {
            color <- indics[which(indics$name == indic), "color"]
            aggreg_indic <- aggreg.indic.P(indic_SeSp, indic, 0, 1)
            lines(aggreg_indic$P, aggreg_indic$moy, type="l",
286               col=color,
                  lty=indics[which(indics$name == indic), "line_type"],
                  lwd=2)
            lines(aggreg_indic$P, aggreg_indic$quant025, type="l",
                  col=color, lty=3)
291         lines(aggreg_indic$P, aggreg_indic$quant975, type="l",
                  col=color, lty=3)
        }

        legend("topright", legend=text.legends, col=indics$color,
296            lty=indics$line_type, lwd=2, inset=c(-0.2, 0),
               box.lty=0, cex=.8, seg.len = 4)
    }
}

301
## In case of "inversion"

indic_shift <- seq(from=0,length.out=nrow(indics),by=0.005)
indic_shift <- indic_shift - median(indic_shift)
306
for (Se in list_Se) {
    for (Sp in list_Sp) {
        indic_SeSp <- rbind(indicators[which(indicators$Se == Se
                                             & indicators$Sp == Sp
311                                          & indicators$P == 0.1),],
                            indicators[which(indicators$Se == Sp
                                             & indicators$Sp == Se
                                             & indicators$P == 0.9),])

316     par (mar=c(5,4,4,6)+.1, xpd=T, las=1, bty="n")

        y_min <- min(c(0, indic_SeSp$AUCnorm, indic_SeSp$Kappa,
                       indic_SeSp$Co, indic_SeSp$Fmes, indic_SeSp$phi))

321     plot(c(0,1),
             c(y_min,1),
             type="n",
             main=paste("Se = ", Se, ", Sp = ", Sp, sep=""),
```

```
             xlab="situation",
             ylab="indicators",
             xaxt="n"
             )

      axis(1,at=c(0.1,0.9),labels=c("normal", "inversion"))

      for (indic in  indics$name) {
          color <- indics[which(indics$name == indic), "color"]
          aggreg_indic <- aggreg.indic.P(indic_SeSp, indic, 0, 0)
          lines(aggreg_indic$P + indic_shift[which(indics$name == indic)],
                aggreg_indic$moy,
                type="l",
                lty=indics[which(indics$name == indic), "line_type"],
                lwd=2,
                col=color)

          for (Prev in aggreg_indic$P) {
              lines(rep(Prev + indic_shift[which(indics$name == indic)], 2),
                    as.numeric(aggreg_indic[which(aggreg_indic$P == Prev),
                                            c("quant025","quant975")]),
                    col=color,
                    lwd=2)
          }
      }

      legend("topright", legend=text.legends, col=indics$color,
             lty=indics$line_type, lwd=2, inset=c(-0.2, 0),
             box.lty=0, cex=.8, seg.len = 4)
  }
}
```

Annexe B

Ensemble des résultats graphiques

Cette section contient tous les résultats graphiques confrontant les variations des indicateurs dans les trois situations étudiées :

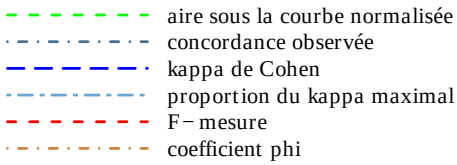
— prise en compte de l'ensemble des taux de prévalence possibles (figure B.1),

— restriction aux taux de prévalence faibles, inférieurs ou égaux à 0,1 (figure B.2),

— étude des deux situations inversées (figure B.3).

Les graphiques sont représentés dans une matrice permettant de comparer toutes les combinaison de Se et Sp.

La légende des couleurs et des types de ligne est la suivante :



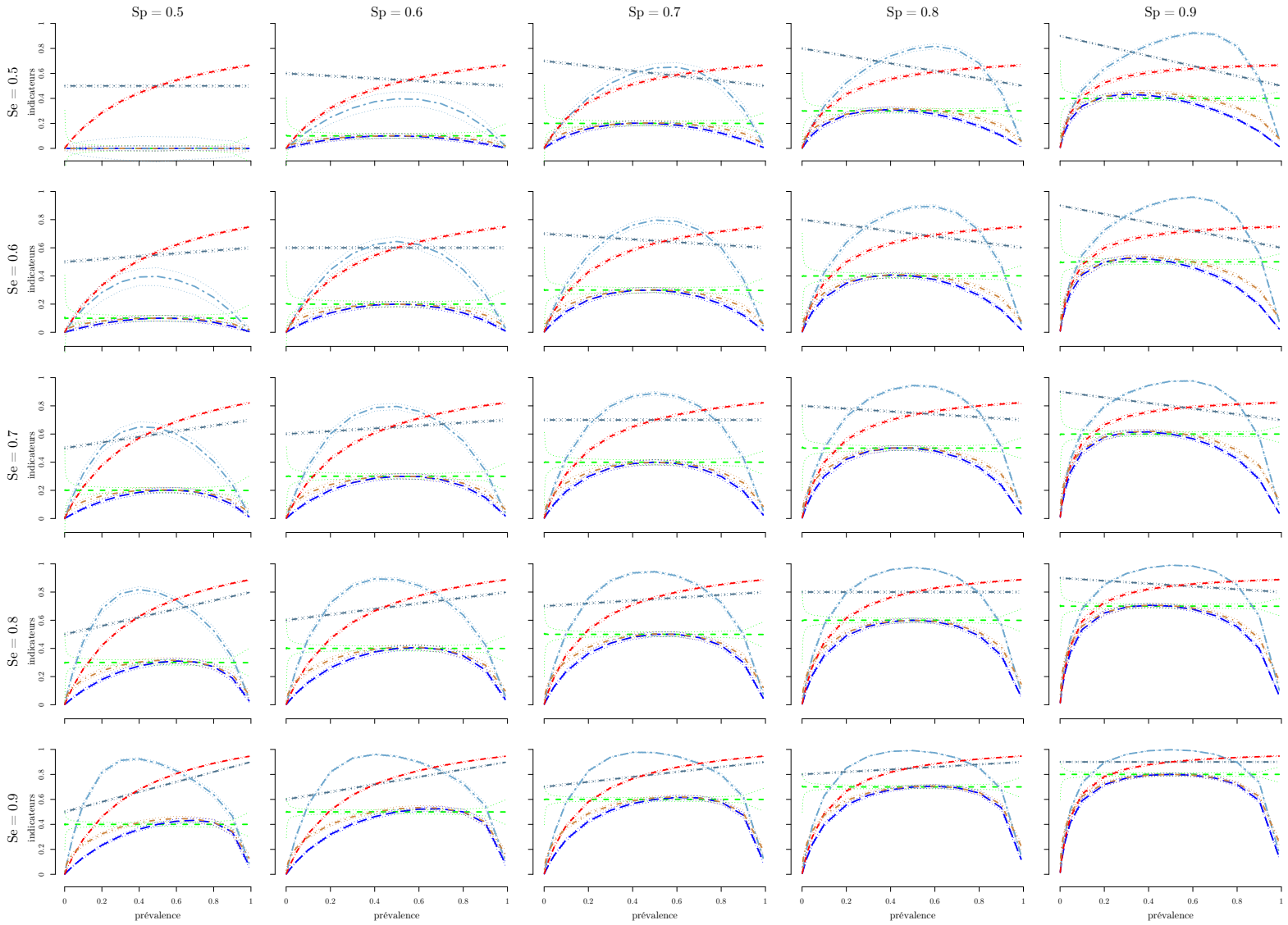


FIGURE B.1 – Variations des indicateurs en fonction du taux de prévalence, pour différents couple de sensibilité (Se) et spécificité (Sp)

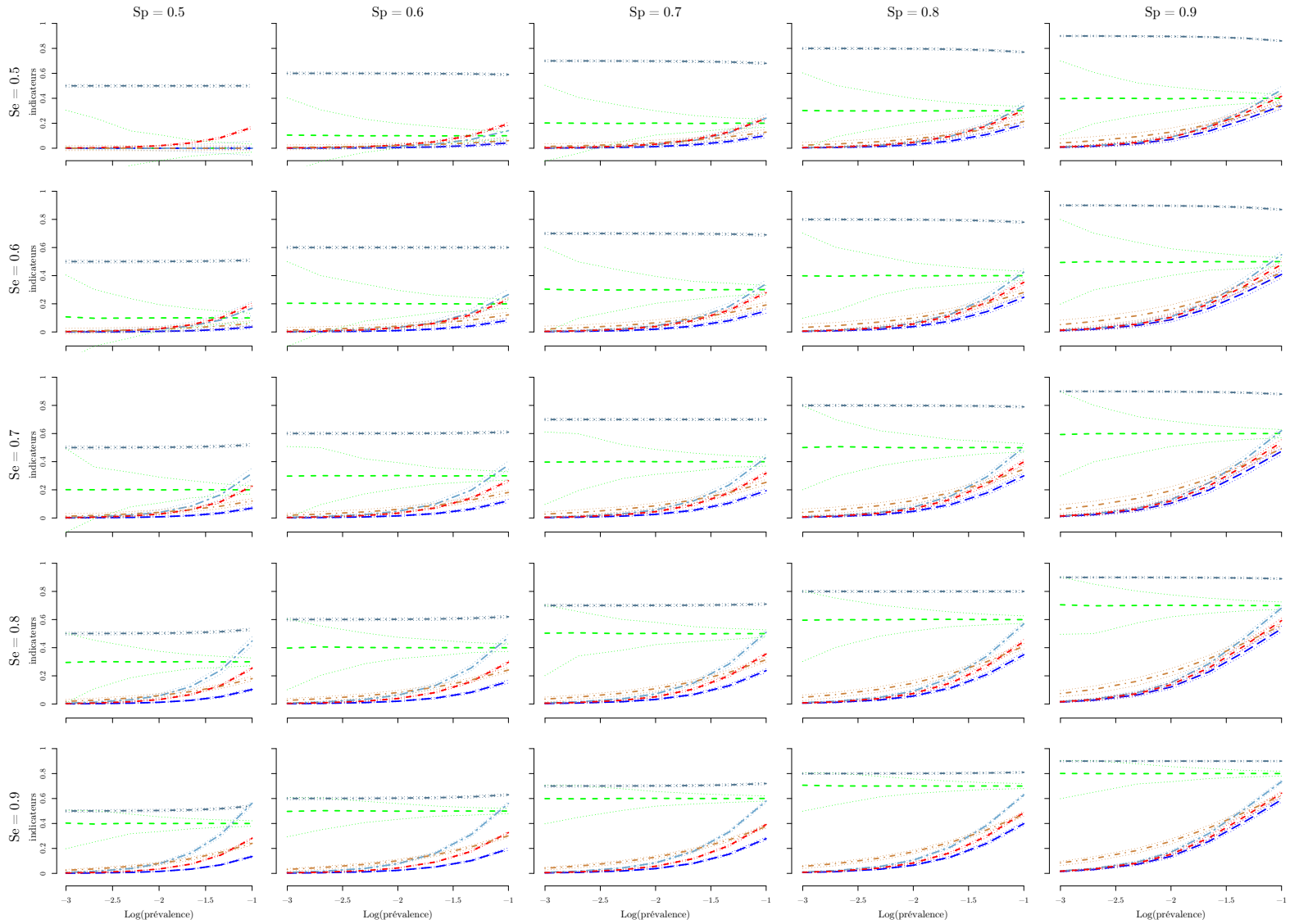


FIGURE B.2 – Variations des indicateurs en fonction du logarithme du taux de prévalence, pour différents couple de sensibilité (Se) et spécificité (Sp) et en se limitant aux taux de prévalence faibles ($\leq 0{,}1$)

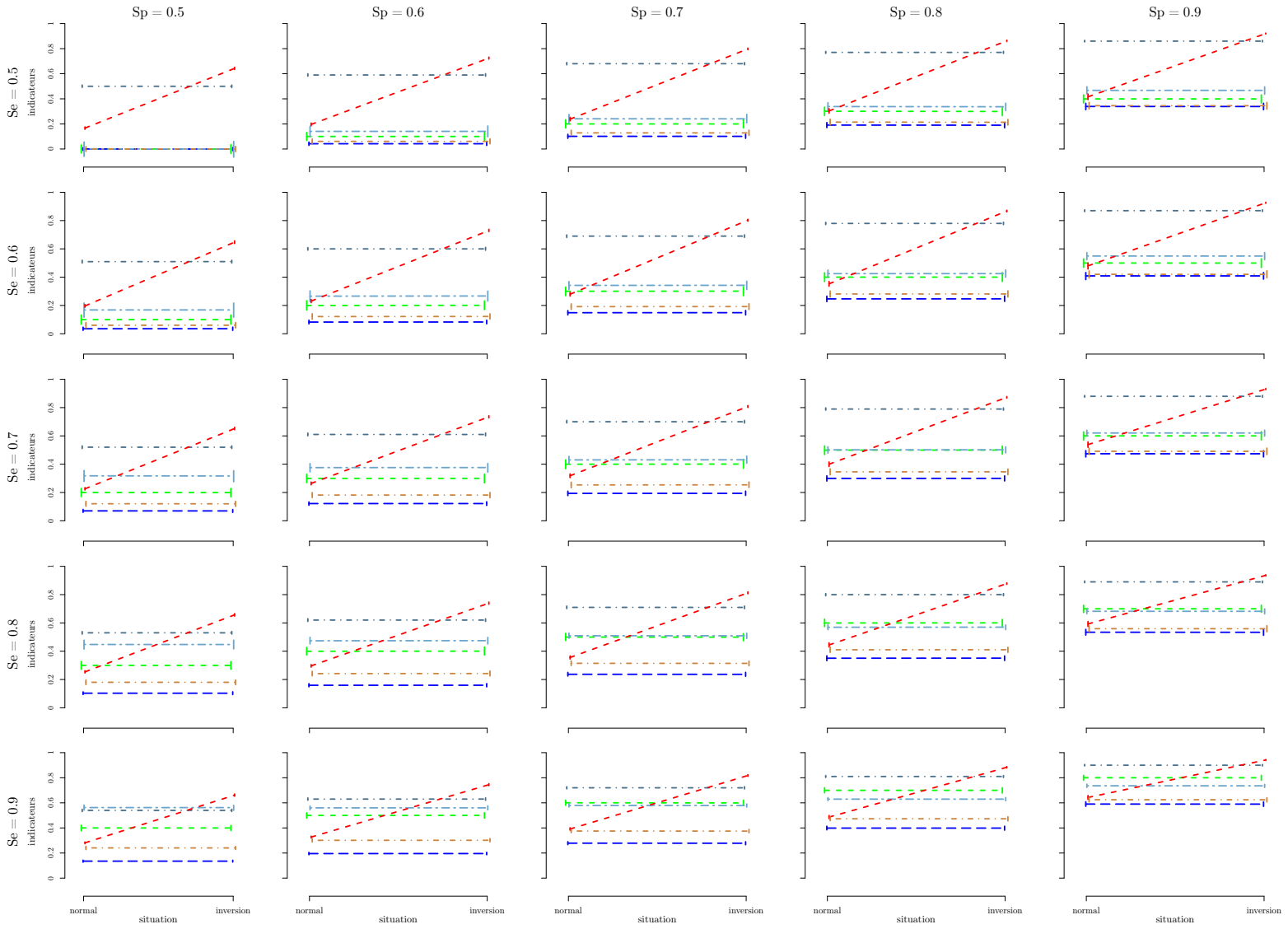


FIGURE B.3 – Variations des indicateurs selon deux situations opposées, « normale » (taux de prévalence = 0.1, sensibilité = Se, spécificité = Sp) et « inversée » (taux de prévalence = 0.9, sensibilité = Sp et spécificité = Se)

Annexe C

Informations sur ce travail

C.1 Matériel

Ce travail a été réalisé exclusivement à l'aide de logiciels libres :

— la composition de ce document a fait appel à LaTeX [1] et la classe yathesis [2] destinée aux thèses selon les spécifications du Ministère de l'Éducation nationale, de l'Enseignement supérieur et de la Recherche [3] ; l'éditeur de texte utilisé est le logiciel Emacs [4] et son extension AUCTeX [5] ,

— les analyses statistiques ont été réalisées avec le logiciel R version 3.2.2 [29], utilisé à travers le logiciel RStudio [6] ,

— le tout sur une plate-forme Linux, distribution Debian 8.5 [7] pour processeur 64 bits.

C.2 Licence

Cette thèse est destinée à une diffusion large.

Néanmoins, parce qu'elle sera publiée dans une revue, elle est soumise à des restrictions de diffusion pendant une période d'« embargo », généralement de six

1. https://www.latex-project.org/
2. https://www.ctan.org/pkg/yathesis
3. Ministère de l'enseignement supérieur et de la recherche, éd. Guide pour la rédaction et la présentation des thèses. À l'usage des doctorants. 2007. url : http://www.u-bordeaux1.fr/fileadmin/images-PDF/DOCUMENTATION/mylene/guidoct.pdf
4. https://www.gnu.org/software/emacs/
5. https://www.gnu.org/software/auctex/
6. https://www.rstudio.com/
7. https://www.debian.org/

mois à un an (selon les directives du journal). Passé ce délai, l'ensemble du travail (script R, graphiques, sources LaTeX, ce document) sera disponible sous licence Creative Commons Attribution 4.0 International [8] sur le compte GitHub de l'auteur (`https://github.com/s-degoul`).

8. `http://creativecommons.org/licenses/by/4.0/`

Auteur : **Nom :** Degoul **Prénom :** Samuel

Date de soutenance : 16 septembre 2016

Titre de la thèse : Comment mesurer la performance d'un test diagnostique ? Présentation et comparaison d'indicateurs.

Thèse – Médecine – Lille 2016

Cadre de classement : Santé publique

DES + spécialité : Anesthésie-Réanimation

Mots-clés : test diagnostique, indicateur de performance, aire sous la courbe, kappa de Cohen, F-mesure, coefficient phi

Résumé :

Contexte. Ce travail vise à comparer le comportement statistique d'indicateurs synthétiques de performance d'un test diagnostique binaire pour aider le chercheur dans le choix de l'indicateur le mieux adapté pour son étude.

Méthode. Les indicateurs étudiés sont l'aire sous la courbe ROC, le coefficient kappa de Cohen et la proportion de sa valeur maximale, la F-mesure, la concordance observée et le coefficient phi. Au moyen de simulations informatiques basées sur plusieurs valeurs de sensibilité et spécificité du test et de taux de prévalence de l'état à diagnostiquer, les valeurs moyennes et les intervalles de fluctuation de ces indicateurs ont été comparés graphiquement. Différentes situations particulières mais rencontrées en pratique ont été évaluées.

Résultats. Les coefficients kappa et phi se montrent symétriques par rapport aux variations de taux de prévalence, avec une valeur maximale quand les échantillons sont équilibrés. Au contraire, la F-mesure semble anormalement optimiste quand le taux de prévalence est élevé. Elle est de plus le seul des indicateurs étudiés sensible à l'inversion des états « pathologique » et « sain ». La concordance observée surestime systématiquement la performance du test. Il en est de même pour la proportion du kappa maximal, sauf lorsque le taux de prévalence est faible. Enfin, l'aire sous la courbe montre de larges fluctuations de l'intervalle de variation pour des taux de prévalence extrêmes, à l'opposé des autres indicateurs.

Conclusion. Les coefficients kappa et phi semblent être les indicateurs les plus à même de refléter, dans les situations étudiées, la performance d'un test diagnostique en terme d'utilisabilité par les cliniciens. Néanmoins, la valeur d'un indicateur seule ne saurait contenir toute l'information sur cette performance.

Composition du Jury :

Président :

Monsieur le Professeur Gilles Lebuffe

Assesseurs :

Monsieur le Professeur Régis Beuscart
Monsieur le Professeur Alain Duhamel
Monsieur le Docteur Jean-Marie Renard
Monsieur le Docteur Emmanuel Chazard