



## SOUS LA LOUPE

Janvier 2023



## DÉMYSTIFIER LES ROC ET LES AUC

Par MATHIEU COUTURE PH.D. Psychologie  
Responsable du développement des pratiques | RIMAS

Pourquoi tenter cet exercice de vulgarisation de concepts aussi peu intéressants que les courbes ROC (*Receiver Operating Characteristic*; voir plus loin) et l'aire sous la courbe (AUC = *area under curve*)?!

La réponse est simple et angoissante à la fois : toutes les praticien.ne.s utilisant des outils d'évaluation du risque de récidive doivent comprendre *sommairement* ces concepts afin de bien sélectionner les outils qu'ils.elles utilisent, bien saisir les limites de ces outils, mieux nuancer leurs conclusions dans leurs rapports d'évaluation et savoir défendre l'utilisation de ces outils en Cour.

Je vous rassure. Il n'est pas nécessaire de comprendre tous les tenants et aboutissants de ces concepts. Nous ne sommes pas des statisticien.ne.s. Cependant, nous utilisons des tests et devons être aptes à représenter adéquatement ces tests. Ceci requiert un niveau de connaissances, disons, intermédiaire, qui implique entre autres la compréhension des notions de courbes ROC et d'AUC.

### ÉTAPE 1 : DISTINGUER UN « VRAI POSITIF » ET UN « FAUX POSITIF »



Le « vrai positif » est l'individu à qui on administre un test et qui « possède » réellement la caractéristique mesurée par le test. Par exemple, une personne passe un test COVID à la maison. Le test est positif. Des analyses en laboratoire confirment ensuite la présence du virus. Le score initial au test-maison est donc un « vrai positif ».



Le « faux positif » est l'individu à qui on administre un test et qui « ne possède » finalement pas la caractéristique mesurée par le test. À ce même test COVID déclaré positif à la maison, il obtient des analyses en laboratoire qui s'avèrent négatives. Son score initial au test-maison était donc un « faux positif ».

Pour le sujet qui nous intéresse, le risque de récidive, c'est plus compliqué. L'outil d'évaluation du risque de récidive n'est pas dichotomique comme un test COVID. Il ne dira pas : « OUI, cet individu récidivera » ou « NON, il ne récidivera pas ». Il offrira plutôt un degré de risque de récidive.

Afin de savoir s'il s'agit d'un « vrai positif », il faut que cet individu qui récidive après un temps donné ait obtenu un score plus élevé qu'un individu similaire n'ayant pas récidivé. Mais ce n'est pas si simple. Étant donné qu'il y a un degré d'imprécision important, c'est un *effet de groupe* que l'on recherche.

Ainsi, il faut que dans l'ensemble, les individus qui récidivent obtiennent *généralement* un score plus élevé au test (vrais positifs) que les individus qui ne récidivent pas. Le taux de « faux positifs » devrait être limité, c'est-à-dire que les scores des non-récidivistes devraient *généralement* être plus faibles que ceux des récidivistes (et être alors des « vrais négatifs »).

Dans notre contexte, évaluer si un individu a récidivé ou non, au cours d'une période donnée, constitue l'équivalent des analyses en laboratoire pour confirmer un test COVID. Des périodes entre 3 et 20 ans sont considérées par les études sur les outils d'évaluation du risque afin de constater s'il y a eu récidive ou non.

## ÉTAPE 2 : AVOIR UNE PETITE IDÉE DE CE QU'EST UNE COURBE ROC

« Une courbe ROC (Receiver Operating Characteristic) est un graphique qui montre les performances d'un modèle de classification à tous les seuils de classification. Cette courbe représente deux paramètres:



Taux de *vrais positifs*



Taux de *faux positifs* »

LIEN : [Classification: courbe ROC et AUC | Machine Learning | Google Developers](#)

« HEIN?! »...

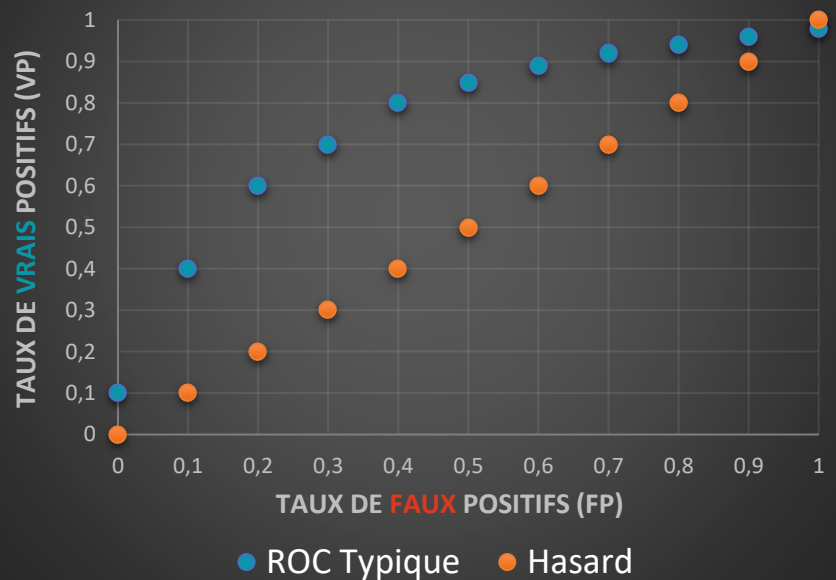
Au fond, c'est comme si on choisissait un score à l'outil pour dire : à partir de « ce » score, on regarde combien on obtient de « vrais positifs » et combien on obtient de « faux positifs ».

Par exemple, si on choisit le score « 1 », comme seuil à la Statique-99R, on se dit que les personnes qui ont « 1 et + » seront récidivistes, alors que les personnes qui ont « moins de 1 » ne seront pas récidivistes. On regarde alors combien de personnes classées « récidivistes » avec ce seuil de décision ont réellement récidivé au cours d'une période donnée (vrais positifs) et combien n'ont pas récidivé (faux positifs).

Plutôt que de calculer chaque combinaison Vrais positifs/Faux positifs pour chaque score choisi comme seuil, un algorithme (qui dépasse l'objectif du présent texte) calcule plutôt *l'aire sous la courbe* afin de fournir une idée de la capacité prédictive de l'outil.

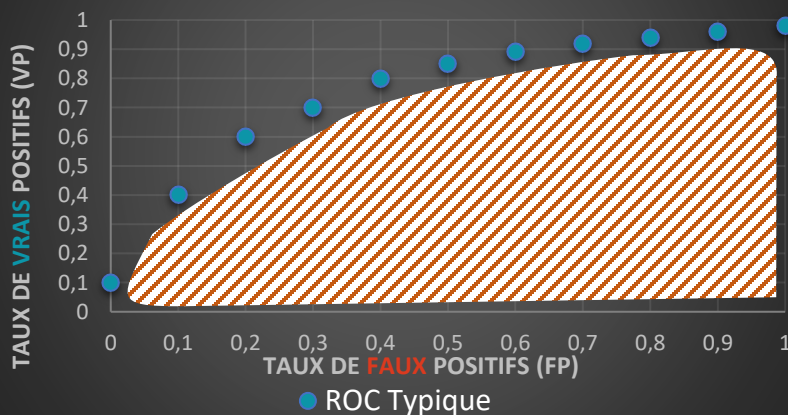
### GRAPHIQUE SIMPLIFIÉ D'UNE COURBE ROC

Taux de FP et de VP selon différents seuils de classification



### GRAPHIQUE SIMPLIFIÉ D'UNE COURBE ROC

Taux de FP et de VP selon différents seuils de classification



## ÉTAPE 3 : AVOIR UNE PETITE IDÉE DE CE QU'EST UNE AIRE SOUS LA COURBE (AUC)

L'aire sous la courbe (*area under curve*), c'est la portion géométrique se situant sous la courbe ROC (aire sous la courbe ROC).

Cet « espace » est quantifié par un algorithme que nous n'avons pas besoin de comprendre (fiou!).

*Le AUC, c'est donc ce calcul de l'aire sous la courbe, entre le point (0,0) et le point (1,1) du graphique.*

« L'AUC fournit une mesure globale des performances pour tous les seuils de classification possibles. L'une des façons d'interpréter l'AUC est la probabilité que le modèle classe un exemple positif aléatoire plus particulièrement qu'un exemple négatif aléatoire ».

LIEN : [Classification: courbe ROC et AUC | Machine Learning | Google Developers](#)

« BEN KIN! »...

Ok. Imaginons deux récipients. Un récipient contient les scores de personnes qui ont récidivé après, disons, un suivi de 5 ans (récidivistes). L'autre contient les scores de personnes qui n'ont pas récidivé (non-récidivistes). On pige au hasard un score dans chaque récipient. Ce qu'on souhaite, c'est que le score du récidiviste soit plus élevé que celui du non-récidiviste.

Supposons qu'on répète le processus des centaines de fois, on souhaite que, *généralement*, les scores des récidivistes soient plus élevés que les scores des non-récidivistes. Si ce n'est pas le cas, on remet en doute la validité de l'outil.

L'aire sous la courbe donnera un nombre compris entre 0 et 1, qui exprime *la probabilité que le score d'un récidiviste soit supérieur à celui d'un non-récidiviste*. Dans un monde parfait, cette probabilité serait de 1 (100 %). En pratique, elle tourne généralement autour de 0,70 pour les outils d'évaluation du risque actuariels. Ainsi, il y a 70 % de chance qu'en pigeant dans les deux récipients, le score du récidiviste soit plus élevé que celui du non-récidiviste.

### VOICI LES BARÈMES AFIN DE CATÉGORISER UN SCORE AUC :

**0,5 = Hasard (aucune valeur prédictive).** Il y a autant de chances qu'un récidiviste ait un score supérieur ou inférieur au non-récidiviste. Faire un « pile ou face », plutôt que d'utiliser cet outil, aurait le même niveau d'efficacité. L'outil est inopérant. (Équivalent avec l'indice de la taille d'effet,  $d$  de Cohen = 0,20).

**0,64 = Valeur prédictive modérée.** La probabilité que le score du récidiviste soit supérieur à celui du non-récidiviste est de 64 %. C'est « modéré » en considérant la complexité des prédictions du comportement criminel, dans l'état actuel de nos connaissances. D'autres diraient que c'est une valeur très faible, considérant les impacts possibles de notre classification sur l'individu concerné (Équivalent avec l'indice de la taille d'effet,  $d$  de Cohen = 0,50).

**0,71+ = Valeur prédictive élevée.** La probabilité que le score du récidiviste soit supérieur à celui du non-récidiviste est de 71 % ou plus. C'est « élevé » en considérant la complexité des prédictions du comportement criminel, dans l'état actuel de nos connaissances. D'autres diraient que c'est encore trop faible (Équivalent avec l'indice de la taille d'effet,  $d$  de Cohen = 0,80).

**0,4 et moins (environ) = Valeur prédictive inverse.** Les non-récidivistes obtiennent généralement un score plus élevé que les récidivistes. La cause de ce résultat peut varier et mérite d'être analysée.

Si plusieurs études concernant un outil fournissent systématiquement des AUC sous les 0,60 (60 %), on peut douter de sa valeur prédictive et commencer à penser sérieusement à écarter cet outil de nos évaluations. Le AUC n'est cependant pas le seul critère à considérer. Il faut examiner, entre autres, les dimensions évaluées par l'outil, la valeur de l'échantillon de développement et plusieurs autres paramètres importants.

### ÉTAPE 4 : CONNAÎTRE LES LIMITES DU AUC ET CE QUI L'INFLUENCE

Un outil qui offre une probabilité d'identifier un individu qui récidivera à 70 %, c'est quand même plutôt intéressant. N'empêche, il demeure 30 % de chances que cet individu ne récidive pas, malgré un score qui dépasse le seuil donné (on comprend ici d'autant plus l'importance de bénéficier d'outils qui proposent de tels seuils). Certains préfèrent abandonner l'évaluation du risque, ce qui peut être une réaction excessive; d'autres considèrent que cette probabilité est significative et que des améliorations sont encore possibles.

Il serait envisageable, en théorie, d'améliorer la valeur prédictive des outils en jouant sur les paramètres qui influencent le AUC. Par exemple, en identifiant les items les plus associés au risque de récidive ou en augmentant la grandeur des échantillons de validation. Certains outils récents fournissent parfois des AUC flirtant avec le 0,80 (80 %), ce qui est vraiment bien (p.ex., certaines études utilisant l'outil du VRS:SO).

Voici quelques facteurs qui peuvent influencer l'ampleur du AUC :

1. **Le type de récidive évalué** (récidive sexuelle, récidive violente, récidive générale) : Souvent, un outil cible mieux un type de récidive particulier. Certains outils, peut-être plus complets, arrivent parfois à prédire similairement différents types de récidive. Il peut arriver qu'un outil d'évaluation du risque de récidive sexuelle prédise paradoxalement mieux la récidive violente ou vice-versa. Il est important de savoir quel outil évalue mieux quel type de récidive.
2. **Le score utilisé** (score global, score à une sous-échelle, jugement final sur l'outil, etc.) : Certains scores fonctionnent mieux que d'autres, selon l'outil utilisé.
3. **La sous-population étudiée** (caucasiens vs minorités ethniques, population avec difficultés sur le plan de la santé mentale vs non, etc.) : Certains outils ont une valeur prédictive simplement meilleure avec des populations spécifiques.
4. **La qualité de la cotation** : les évaluateurs expérimentés et ceux qui procèdent à des cotations comparées avec des collègues obtiennent de meilleurs indices de fidélité interjuges et, ainsi, des valeurs prédictives (AUC) généralement plus élevées.
5. **Des items manquants** : trop d'items manquants réduit la valeur prédictive des outils. Il est pertinent de connaître les limites de chaque outil en ce sens. Certains items sont plus importants que d'autres pour assurer la valeur prédictive d'un outil.
6. **Des échantillons de validation trop petits** : un taux de récidivistes trop faible réduira la valeur prédictive de l'outil. Plus l'échantillon est petit, plus le nombre de récidivistes risque d'être faible. Les études de validation doivent idéalement être réalisées avec de grandes cohortes (d'où le recours fréquent aux méta-analyses afin d'évaluer la capacité prédictive d'un outil).
7. Etc.

## CONCLUSION

En considérant tous les facteurs qui influencent la valeur prédictive d'un outil, on se dit que même un outil ayant une valeur prédictive élevée n'est pas nécessairement approprié pour un contexte d'évaluation donnée. Il faut être à l'affût des forces et des faiblesses de l'outil utilisé et des caractéristiques de l'individu rencontré afin de mieux choisir.

Très prochainement, deux grilles répertoriant l'ensemble des outils utilisés auprès des auteur.e.s d'infraction à caractère sexuel vous seront transmises (grille Adultes et grille Adolescents). Dans celles-ci, chaque outil est détaillé et un accent est mis sur les capacités prédictives des outils. Ces grilles vous aideront à faire des choix d'outils éclairés.

Si vous avez des interrogations et n'avez pas tout compris, bienvenus dans mon club! N'hésitez pas à me contacter, il me fera plaisir de discuter et d'essayer de démêler des choses avec vous!



*Je vous invite aussi à poser vos questions sur le forum [Facebook du RIMAS](#), afin que toutes puissent bénéficier de l'approfondissement issu des réponses que nous pourrions trouver ensemble à ces questions!*