Des probabilités et des cotes

Philippe Marchal

Quel niveau de difficulté conceptuelle représente le fait de distinguer p et $p/(1-p)$? Lorsqu'on est habitué à l'usage des mathématiques, on pourrait penser que toute personne ayant suivi un cursus scientifique au lycée, et *a fortiori* toute personne ayant suivi des études scientifiques, devrait n'avoir absolument *aucune* difficulté à ce propos. Il n'en est rien, nous l'allons montrer tout à l'heure.

Commençons par citer un chercheur en médecine à propos de cette subtile distinction : "C'est tellement complexe que malheureusement (...) Il y a aussi un usage pour nous d'utiliser certaines métriques, certaines mesures en fait par simplification. C'est voulu, on sait que ça distord un peu mais c'est l'usage consacré. On pourrait refaire avec les bonnes choses. Ça complexifierait beaucoup la méthode (...)" [1]. Ce chercheur a au moins le mérite de la franchise : pour lui, distinguer p et $p/(1-p)$ relève d'un niveau de complexité qui le dépasse.

Des chances et des paris

On ne va pas ici reprendre toute la théorie des probabilités mais rappeler brièvement ceci :

– la probabilité qu'un événement se produise est un nombre p compris entre 0 et 1,

– à une probabilité p on peut associer une cote ("odds" en anglais) qui est donnée par $p/(1-p)$. [2]

Le terme "cote" renvoie au vocabulaire des paris. Supposons que vous misiez sur la victoire d'un cheval à c contre 1. Cela veut dire que

– si le cheval perd, vous perdez votre mise,

– s'il gagne, on vous rembourse votre mise et en sus, on vous donnex c fois votre mise.

Supposons qu'on connaisse la probabilité p que le cheval perde. À quelle condition le jeu sera-t-il équitable ? L'espérance de gain se calcule simplement comme suit : avec probabilité p vous perdez votre mise, qu'on note m ; avec probabilité $1-p$ vous gagnez votre mise multipliée par la cote, soit cm. L'espérance de gain est alors

$$E = -pm + (1-p)cm$$

Le jeu est équitable si l'espérance de gain est nulle, autrement dit si $-pm + (1-p)cm = 0$, soit

$$c = p/(1-p)$$

La cote représente donc le facteur multiplicatif de gain pour un pari équitable. Bien sûr, les maisons de jeu n'ont pas intérêt à proposer de paris équitables : leur but est de gagner de l'argent mais ceci est une autre question.

1. voir à 1h10 sur la vidéo `https://www.youtube.com/watch?v=xm5GvREYQMY`
2. si $p = 1$, la cote est infinie ; il s'agit d'un cas limite sur lequel on ne va pas s'attarder ici.

Si maintenant on a deux probabilités p et q, le rapport de ces probabilités, aussi appelé risque relatif, est p/q ; le rapport des cotes, "odds ratio" en anglais, est

$$\frac{\left(\frac{p}{1-p}\right)}{\left(\frac{q}{1-q}\right)}$$

Naturellement si $p \neq q$,

$$\frac{p}{q} \neq \frac{\left(\frac{p}{1-p}\right)}{\left(\frac{q}{1-q}\right)}$$

Des risques et des cotes

L'article "odds ratio" de wikipedia anglophone nous apprend que selon une étude datée de 2001, environ 1/4 des articles publiés dans certaines revues de médecine confondaient les notions de risque relatif et de rapport des cotes. Un contributeur de wikipedia qualifie ceux qui commettent cette erreur du terme peu élogieux de "uncomprehending authors", qu'on pourrait traduire un peu familièrement par "auteurs malcomprenants".

Plus récemment, on retrouvait cette erreur dans un article scientifique censé évaluer la surmortalité due à l'usage de l'hydroxychloroquine pour traiter la covid-19 [3]. Interrogé sur cette bourde, un des auteurs répondait par les propos cités plus haut. Cet article, coécrit par six chercheurs, publié après relecture minutieuse (ou pas) par d'autres chercheurs, a été abondamment relayé par de nombreux scientifiques (dont au moins un, une devrais-je dire, siège à l'Académie des sciences) sans que cette erreur de niveau lycée soit relevée. Elle a été publiquement exposée par le mathématicien Vincent Pavan un peu plus tard [4].

La surmortalité due à l'hydroxychloroquine est calculée dans l'article par la formule

$$\widehat{S} = H \times T \times M \times (Odd - 1)$$

où H est le nombre de patients hospitalisés, T est le taux de patients traités avec l'hydroxychloroquine, M est le taux de mortalité sans usage de l'hydroxychloroquine et Odd est le rapport des cotes. Cette formule est évidemment fausse [5]. De fait cette surmortalité est donnée par

$$S = H \times T \times (M' - M)$$

où M' est le taux de mortalité quand on utilise l'hydroxychloroquine.

Les quantités M et M' sont des probabilités liées entre elles par le rapport des cotes

$$Odd = \frac{\left(\frac{M'}{1-M'}\right)}{\left(\frac{M}{1-M}\right)}$$

Des manipulations algébriques élémentaires permettent d'exprimer M' en fonction de M et de Odd pour obtenir la formule correcte donnant la surmortalité :

$$S = H \times T \times M \times (Odd - 1) \times \frac{1 - M}{1 + M(Odd - 1)}$$

3. "Deaths induced by compassionate use of hydroxychloroquine during the first COVID-19 wave : an estimate", *Biomedicine and Pharmacotherapy*, 2024.

4. J'avais signalé l'erreur quelques jours avant Vincent Pavan sur le forum (semi-public) de mon laboratoire, avec des commentaires peu amènes que la décence m'interdit de reproduire ici.

5. "Demander à un chercheur en médecine une formule mathématique correcte, c'est comme demander à un sumotori de faire du saut à la perche", proverbe japonais

Le calcul serait sans doute considéré comme plutôt difficile pour le brevet des collèges mais plutôt facile au niveau baccalauréat pour un élève ayant suivi l'option mathématiques. Le fait qu'un professeur en médecine considère comme "tellement complexe" la multiplication par le facteur

$$F = \frac{S}{\widehat{S}} = \frac{1-M}{1+M(Odd-1)}$$

pourra, au choix, faire sourire ou laisser mal à l'aise. L'article, reprenant des données d'autres études, considère que M varie suivant les pays, de 0,055 pour la Turquie à 0,228 pour l'Italie, tandis que Odd ne dépend pas des pays [6] et vaut $Odd = 1,11$. En particulier, quel que soit le pays, $1+M(Odd-1) > 1$ d'où $F < 1$. Si on reprend les données de l'article et qu'on refait les calculs corrects, on trouve une surmortalité totale de 13 144 au lieu de 16 990 comme annoncé par les auteurs. Ainsi la célèbre approximation

$$2 + 2 = 5$$

qui est, on le sait, couramment utilisée par certains élèves de CP promis à un brillant avenir de chercheur en médecine, est moins mauvaise que celle de l'article puisqu'elle ne surestime le bon résultat que de 25%, alors que l'article surestime la surmortalité de plus de 29%.

Plus généralement, que la distinction entre ces deux notions de risque relatif et de rapport des cotes soit considérée comme aussi difficile par de nombreux chercheurs n'étonnera que les personnes superstitieuses persuadées que de longues études couronnées par un doctorat garantissent des compétences intellectuelles supérieures à celles d'un bon lycéen.

Des contrôles d'identité

La pratique des contrôles d'identité par la police est un sujet de polémique récurrent et sans surprise, la confusion mentionnée plus haut s'y retrouve. Une étude très médiatisée à sa sortie et encore abondamment citée de nos jours est celle de Jobard et Lévy. Dans un premier rapport [7], les auteurs y expliquaient : "l'odds-ratio est le ratio de 2 probabilités (...)". Autrement dit, sans même parler de compétence mathématique, on constate que les auteurs se montrent incapables de recopier une définition.

Dans un texte ultérieur [8], les auteurs prétendent effectuer une régression logistique et calculer des rapports des cotes. Or leurs données ne leur permettent pas d'effectuer ces calculs. N'importe qui ayant un minimum de compréhension des mathématiques devrait le voir facilement mais bien sûr, cela a complètement échappé à toute une communauté universitaire fort peu à l'aise avec l'algèbre de niveau lycée : l'article a été publié dans une des revues les plus prestigieuses du domaine et n'a jamais suscité la moindre critique méthodologique.

Pour résumer, les auteurs ont procédé en deux étapes :

– Une étape d'étalonnage ("benchmarking" en anglais) analysant la composition démographique (sexe, âge, origine ethnique...) de la population présente en certains lieux par échantillonnage des flux entrants.

– Une étape relevant la composition démographique de tous les individus contrôlés par la police en ces lieux à certaines périodes.

La comparaison des proportions de personnes présentes et de personnes contrôlées permet de calculer le risque relatif, *mais pas* le rapport des cotes. Pour prendre un exemple, supposons

6. On pourrait discuter la validité de cette hypothèse mais ce n'est pas notre préoccupation ici.

7. "Police et minorités visibles : les contrôles d'identité à Paris", Open Society Justice Initiative

8. "Mesurer les discriminations selon l'apparence : une analyse des contrôles d'identité à Paris", *Population*, 2012

que les femmes représentent la moitié des personnes présentes mais un dixième des personnes contrôlées. On en déduit que si N est le nombre de personnes présentes et si k femmes ont été contrôlées, le risque d'être contrôlé pour une femme est $k/(N/2)$. Le nombre d'hommes contrôlés est alors $9k$ puisque les femmes représentent une personne contrôlée sur 10. Le risque d'être contrôlé pour un homme est donc $9k/(N/2)$ et le risque relatif est donné par

$$\frac{k/(N/2)}{9k/(N/2)} = \frac{1}{9}$$

On remarque que si on ne connaît pas N, on ne peut calculer aucun des deux risques d'être contrôlé (pour les femmes et les hommes) mais on peut calculer le rapport de ces risques puisque la quantité N disparaît dans l'expression du risque relatif.

En revanche, calculer le rapport des cotes nécessite de connaître aussi le rapport des probabilités de ne pas être contrôlé. Ce rapport est

$$\frac{1-[k/(N/2)]}{1-[9k/(N/2)]} = \frac{N-18k}{N-2k}$$

et ici, aucune simplification ne permet de terminer le calcul si on n'a pas la valeur de N ; de fait, le résultat *dépend* de N. L'article repose donc sur une fabrication puisqu'il prétend calculer des quantités qu'on ne peut déterminer en l'absence d'une donnée cruciale. Il est vraisemblable que les auteurs ont fait comme si la population totale était égale au nombre d'individus relevés lors de la phase d'étalonnage. Ce faisant, N est fortement sous-estimé, ce qui entraîne que les rapports des cotes, lorsqu'ils sont inférieurs à 1, sont sous-estimés. Dans l'exemple donné ci-dessus, la fonction qui à N associe

$$\frac{N-18k}{N-2k} = 1 - \frac{16k}{N-2k}$$

est en effet croissante en N : augmenter N fait augmenter cette quantité, et donc augmenter le rapport des cotes.

Plus généralement, on peut démontrer de manière analogue que la sous-estimation de N entraîne une sous-estimation du rapport des cotes quand celui-ci est inférieur à 1 et une sur-estimation lorsque celui-ci est suprérieur à 1. Ainsi la sous-estimation de N éloigne le rapport des cotes de la valeur 1. Or cette valeur 1 correspond à $p = q$, cas où on n'a pas de de surre-présentation dans les contrôles d'identité. Si l'objectif est prouver qu'il y a surreprésentation, la sous-estimation de N a pour conséquence d'exagérer l'effet qu'on veut mettre en évidence.

Pour donner une idée de la sous-estimation grossière de N si on le confond avec la population échantillonnée lors de la première phase, prenons le cas de Gare du nord dans sa partie gare de surface (par opposition à sa partie souterraine). La population totale relevée lors de la phase 1 y est de 8008 pour un nombre de contrôles de 123 lors de la phase 2 (voir tableau 7 de l'article). Or les auteurs indiquent qu'il y a eu environ 1,25 contrôle par heure, ce qui implique que les 123 contrôles ont eu lieu sur une centaine d'heures. Et il y a largement plus de 8000 personnes présentes à la Gare du nord sur une centaine d'heures !

On notera au passage qu'au sommet même du système universitaire français, à savoir au Collège de France, le professeur Héri Héran (Ranpataplan) [9] confond dans un de ses cours la

9. `https://www.youtube.com/watch?v=uuyOfi7lKUk`

population totale lors de la deuxième étape (la quantité N) avec la population totale échantillonnée lors de la première étape [10]. Rappelons qu'un professeur au Collège de France n'a qu'une vingtaine d'heures de cours à préparer chaque année. Errare humanum est, perseverare diabolicum : Ranpataplan a réitéré son erreur lors d'une conférence quelques mois plus tard [11]. Le pire est qu'en 2010, notre sommité universitaire, présidant un comité Théodule répondant au doux nom de Comedd, rédigeait un des ces innombrables rapports que personne ne lit mais dans lequel il était déjà question des travaux de Jobard–Lévy [12]. Au bout de dix ans, Ranpataplan n'aura donc toujours pas compris le protocole de l'étude, pourtant simplement explicable en quelques lignes. Ce faisant, l'impossibilité de calculer les rapports des cotes avec ce protocole lui sera passé largement par-dessus la tête.

Une autre étude illustre de manière encore plus caricaturale la confusion mentale autour des notions de risque relatif et de rapport des cotes. Un rapport du Défenseur des droits [13] relevait que 80% des jeunes hommes noirs ou arabes avait fait l'objet d'un contrôle d'identité au cours des 5 années passées, contre 16% pour l'ensemble de la population. Les auteurs en concluaient : "ces profils ont ainsi une probabilité 20 fois plus élevée que les autres d'être contrôlés". Le lecteur attentif aura facilement identifié une nouvelle confusion entre le risque relatif et le rapport des cotes.

À l'époque, cette grossière erreur avait été reprise en boucle par la presse - une recherche sur internet permettra de trouver des dizaines d'articles reprenant l'affirmation du Défenseur des droits sans la moindre forme d'esprit critique. Pis encore, les émeutes suite à la mort du prénommé Nahel ont donné l'occasion à de nombreux commentateurs de ressortir ce chiffre grotesque, toujours sans se poser la moindre question. À une époque où les journalistes se targuent de faire de la vérification factuelle, le fait qu'ils ne soient même pas capables de vérifier le résultat de la divison 80/16 est assez révélateur. Il y a un siècle, un élève incapable d'effectuer cette division - sans calculatrice naturellement - n'aurait jamais eu son certificat d'étude.

Des risques de contamination à la covid-19

Une série d'études sur la covid, baptisées comcor et abondamment relayées en période pandémique, a été menée par un légionnaire d'honneur et son équipe de l'Institut Pasteur. Le premier volet de cette étude avait fait l'objet d'une prépublication en français [14] où il était écrit à propos des odds ratios (OR) : "Les ORs représentent le ratio des risques d'être infecté par le SARS-CoV-2 entre exposés et non exposés pendant la période considérée".

Là encore, les auteurs se montrent incapables de recopier une définition. Le plus remarquable est qu'à l'époque, l'auteur principal, âgé de 59 ans, avait déjà derrière lui une carrière d'épidémiologiste de plusieurs décennies au cours de laquelle il avait été constamment confronté à la notion d'odds ratio qu'il comprenait manifestement de travers. Cela ne l'a nullement empêché d'être nommé au Conseil "scientifique" ni d'être adoubé chevalier.

Le lecteur un peu naïf pourra se demander comment cela est possible. Il faut savoir que l'usage des mathématiques par le chercheur non mathématicien et, disons, non physicien et

10. voir à 1h34 sur la vidéo `https://www.youtube.com/watch?v=W01l6gigGFU`
11. voir à 1h05 sur la vidéo `https://www.youtube.com/watch?v=Q55ljlPhiRI`
12. `https://www.vie-publique.fr/files/rapport/pdf/104000077.pdf`
13. "Relations police / population : le cas des contrôles d'identité", 2017
14. "Etude des facteurs sociodémographiques, comportements et pratiques associés à l'infection par le SARS-CoV-2 (ComCor)", 2020

non informaticien, se borne souvent à entrer des données dans un logiciel de statistiques et à interpréter, parfois de manière incorrecte, ce que le logiciel renvoie. Or dans la méthode de régression logistique couramment employée par de nombreux universitaires, le logiciel calcule un rapport des cotes et non un risque relatif.

En 2019, peu avant la pandémie, un article dans la revue *Nature*[15] mettait en garde contre une mauvaise utilisation des statistiques dans la recherche. Coécrit par trois personnes, il était cosigné par 800 chercheurs. Il émettait des recommandations en affirmant que si celles-ci étaient suivies, "people will spend less time with statistical software, and more time thinking" ("les gens passeront moins de temps sur leur logiciel de statistiques et plus de temps à réfléchir"). La question abordée était celle du seuil de significativité statistiques et non la confusion entre risque relatif et rapport des cotes. Mais le principe de réfléchir plus et de moins utiliser son logiciel de statistiques vaut de manière générale.

Le période du ridicule dans cet usage du logiciel de statistiques sans une once de réflexion est probablement atteint par un autre article sur les risques de contamination à la covid-19[16]. Les auteurs s'y proposent de comparer les risques de contamination selon différents critères.

Table. Participant Characteristics by SARS-CoV-2 Status and Results of Logistic Regression Analyses Regarding SARS-CoV-2 Risk

	SARS-CoV-2 status, No. (%)		Univariable analysis[a]	
Variable	Negative (n = 2170)	Positive (n = 749)	OR (95% CI)	*P* value
Baseline				
Age, median (range), y	43.2 (18-73)	40.6 (18-66)	0.98 (0.97-0.99)	<.001
BMI, median (range)	24.4 (14.3-65.8)	24.3 (15.8-44.6)	1.00 (0.98-1.01)	.62
Sex				
Female	1701 (78.4)	597 (79.7)	0.96 (0.78-1.18)	.69
Male	469 (21.6)	152 (20.3)	[Reference]	

Par exemple, le nombre d'hommes infectés (resp., non infectés) est 152 (resp. 469), ce qui permet de calculer la probabilité d'infection

$$p = \frac{152}{152 + 469}$$

et la cote associée

$$\frac{\frac{152}{152+469}}{\frac{469}{152+469}} = \frac{152}{469}$$

pour un homme. On peut faire le même calcul pour les femmes et on trouve un rapport des cotes égal à 0,96. Cependant dans l'unique tableau de l'article, la première ligne nous apprend que l'âge médian des personnes contaminées (resp., non contaminées) est de 40,6 ans (resp. 43,2 ans). On peut évidemment effectuer le même calcul que plus haut

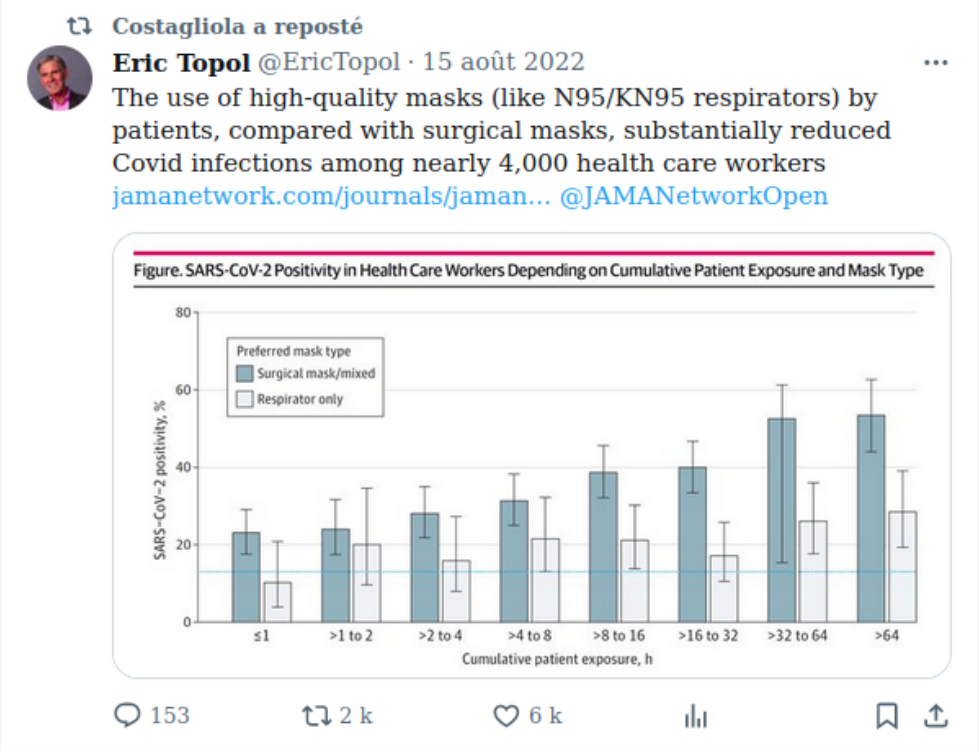
$$\frac{40,6}{43,2}$$

mais la quantité calculée ne s'interprète pas comme une cote ; en fait elle n'a absolument aucun sens statistique. Cela n'a pas empêché les auteurs de faire le calcul et de remplir

15. https://www.nature.com/articles/d41586-019-00857-9
16. Risk of SARS-CoV-2 Acquisition in Health Care Workers According to Cumulative Patient Exposure and Preferred Mask Type, *JAMA Network Open*, 2022

la case correspondante. L'*âge du capitaine* était conçu à l'origine par Flaubert comme une plaisanterie, il est devenu une prophétie. Et un bonheur ne venant jamais seul, les auteurs ont fait le même calcul sans queue ni tête avec l'indice de masse corporelle (BMI en anglais).

Là encore, cet article a été abondamment relayé. Un personnage très suivi sur les réseaux sociaux a ainsi publié un message sur twitter aimé plus de 6000 fois et reposté plus de 2000 fois, notamment par cette académicienne déjà évoquée et que les esprits facétieux appellent madame Charlson [17]



Splendeur de la Science, misère des "scientifiques"

Les exemples passés ici en revue ont eu un écho particulier dans la presse ou sur les réseaux sociaux mais on pourrait en trouver d'autres. Quand le système d'expertise universitaire [18] qui devrait être capable de détecter des erreurs aussi grossières est à ce point défaillant, on s'imagine qu'on n'a vu là qu'une infime partie de de la partie émergée de l'iceberg. Par ailleurs, le présent texte ne mentionne que des problèmes relatifs aux rapports des cotes. Les autres occasions d'erreurs mathématiques dans le monde de la recherche sont innombrables. Tenter de les relever toutes permettrait de toucher du doigt cette notion topologique si pittoresque qu'on appelle *espace totalement inépuisable* [19].

17. Elle avait affirmé publiquement à trois reprises que la proportion d'hospitalisés de moins de 50 ans sans comorbidités lors de la première vague de covid était supérieure à 40%, interprétant de manière fautive l'indice de Charlson : voir `https://twitter.com/DgCostagliola/status/1480088447096401922`, ainsi que ses passages sur France Inter et France Culture autour de cette période. C'est un peu comme si un supposé spécialiste d'aviation disait qu'un airbus vole à 90 km/h.

18. Le rapport du Défenseur des droits était placé sous la supervision scientifique du laboratoire Pacte de Sciences po Grenoble

19. Un espace topologique est inépuisable s'il n'est pas maigre relativement à lui-même. Il est totalement inépuisable si tout sous-espace fermé non vide est inépuisable. Bourbaki, *Topologie générale*

Appendice : un savoureux échange entre modernes Diafoirus

Pr Mathieu Molimard @MathieuMolimard · 2 févr.

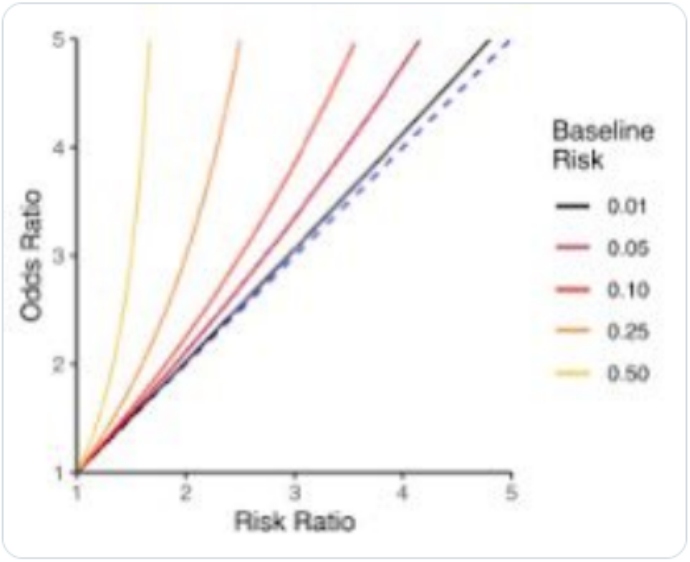
La présentation faite par le “conseil scientifique indépendant” est fallacieuse
Ils se mettent dans des conditions de risque élevé où l'OR ne peut pas estimer le RR (incompétence ou manipulation ?)
L’approximation est correcte et admise en cas de risque faible comme dans l'étude

> **Alexander Samuel** @AlexSamTG · 19 janv.
> En réponse à @AlexSamTG
> Ses chiffres ne remplissent pas les conditions permettant de rapprocher l'OR du RR.
> Car quand p et q sont petits, négligeables devant 1, alors 1-p est proche de 1 et 1-q est aussi proche de 1. OR environ = RR...
> Voir plus

6 12 40 4 k

Jean-Jacques Parienti @JjParienti · 2 févr.

On parle souvent d’un risque inférieur à 10% pour considérer l’approximation acceptable
Au delà on surestime la force d’association avec l’OR



1 105

Pr Mathieu Molimard
@MathieuMolimard

Dans l'étude des 17000 avec un risque de base entre 10 et 20% selon les pays et un OR à 1,1 (qui a aussi un impact) l'erreur est dans l'épaisseur du trait et inférieure à 10%...

9:35 PM · 2 févr. 2024 · **135** vues