95% de confiance, avec une pincée de méfiance

Comment on montre qu'un traitement est efficace

- On part de l'equipoise: on ne sait pas si le ttt est efficace ou pas
- On fait un essai clinique randomisé (ou deux)
- Si le bénéfice du nouveau traitement est « statistiquement significatif », on approuve le nouveau médicament

- Approche critiquée actuellement (ASA, Bayesiens, etc)
- Réponse binaire, le degré d'efficacité n'est pas crucial pour le test
- Risque d'erreur (type 1) est valable asymptotiquement, mais on ne fait qu'un seul essai!!

Neyman J, Pearson E. On the problem of the most efficient tests of statistical hypotheses. Philos Trans R Soc Lond A. 1933;231:289–337.

But we may look at the purpose of tests from another view-point. Without hoping to know whether each separate hypothesis is true or false, we may search for rules to govern our behaviour with regard to them, in following which we insure that, in the long run of experience, we shall not be too often wrong. Here, for example, would be such a "rule of behaviour": to decide whether a hypothesis, H, of a given type be rejected or not, calculate a specified character, x, of the observed facts; if $x > x_0$ reject H, if $x \leqq x_0$ accept H. Such a rule tells us nothing as to whether in a particular case H is true when $x \leqq x_0$ or false when $x > x_0$.

Alternative: estimation

- On formule une question scientifique qui concerne un paramètre (ex: réduction de mortalité grâce au nouveau traitement)
- On fait une étude qui produit un estimateur de ce paramètre
- Ce « point estimate » est la valeur la plus plausible du paramètre au vu des données
- D'autres valeurs du paramètre sont également compatibles avec les données; elles sont représentées par l'intervalle de confiance (IC)
- L'incertitude quant au vrai effet est mise en valeur par l'IC
- A noter le parallèle avec le test:
  - Si on répète l'étude de nombreuses fois, 95% des IC à 95% contiennent la vraie valeur du paramètre
  - 95% est une propriété au long cours de la procédure; pour un résultat donné on ne sait pas s'il contient la vraie valeur ou non

On the Two Different Aspects of the Representative Method: The Method of Stratified Sampling and the Method of Purposive Selection
Author(s): Jerzy Neyman
Source: *Journal of the Royal Statistical Society*, Vol. 97, No. 4 (1934), pp. 558-625

Denote now by $\varphi(\theta)$ the unknown probability distribution *a priori* of θ. Suppose that the general conditions of sampling and the properties of the collective characters θ and x define certain values which these characters may possess. In the example I mentioned above, θ, the proportion of individuals of the given type in the population may be any number between 0 and 1. On the other hand, x, the proportion of these individuals in the sample, say of n, could have values of the form k/n, k being an integer $0 \leq k \leq n$.

The new form of the problem of estimation of the collective character θ may be stated as follows: given any positive number $\varepsilon < 1$, to associate with any possible value of x an interval

$$\theta_1(x) < \theta_2(x) \quad . \quad . \quad . \quad . \quad . \quad . \quad (1)$$

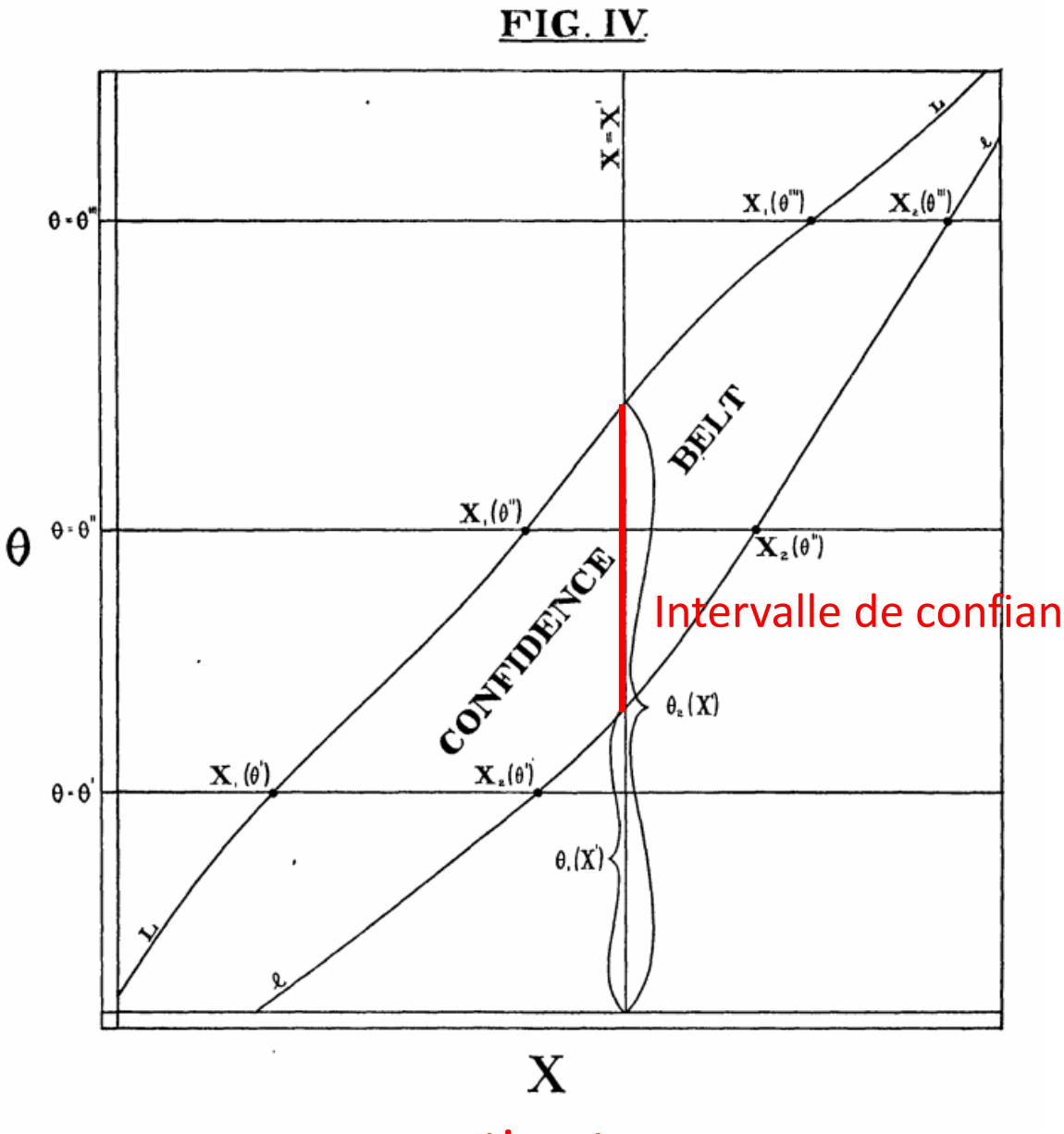
such that if we accept the rule of stating that the unknown value of the collective character θ is contained within the limits

$$\theta_1(x') \leq \theta \leq \theta_2(x'). \quad . \quad . \quad . \quad . \quad . \quad (2)$$

every time the actual sampling provides us with the value $x = x'$, the probability of our being wrong is less than or at most equal to $1 - \varepsilon$, and this whatever the probability law *a priori*, $\varphi(\theta)$.

The value of ε, chosen in a quite arbitrary manner, I propose to call the "confidence coefficient." If we choose, for instance, $\varepsilon = \cdot 99$ and find for every possible x the intervals $[\theta_1(x), \theta_2(x)]$ having the properties defined, we could roughly describe the position by saying that we have 99 per cent. confidence in the fact that θ is contained between $\theta_1(x)$ and $\theta_2(x)$. The numbers $\theta_1(x)$ and $\theta_2(x)$ are what R. A. Fisher calls the fiducial limits of θ. Since the word "fiducial" has been associated with the concept of "fiducial probability" which has caused the misunderstandings I have already referred to, and which in reality cannot be distinguished from the ordinary concept of probability, I prefer to avoid the term and call the intervals $[\theta_1(x), \theta_2(x)]$ the confidence intervals, corresponding to the confidence coefficient ε.

Note: Neyman se fiche complètement du point estimate!!



paramètre

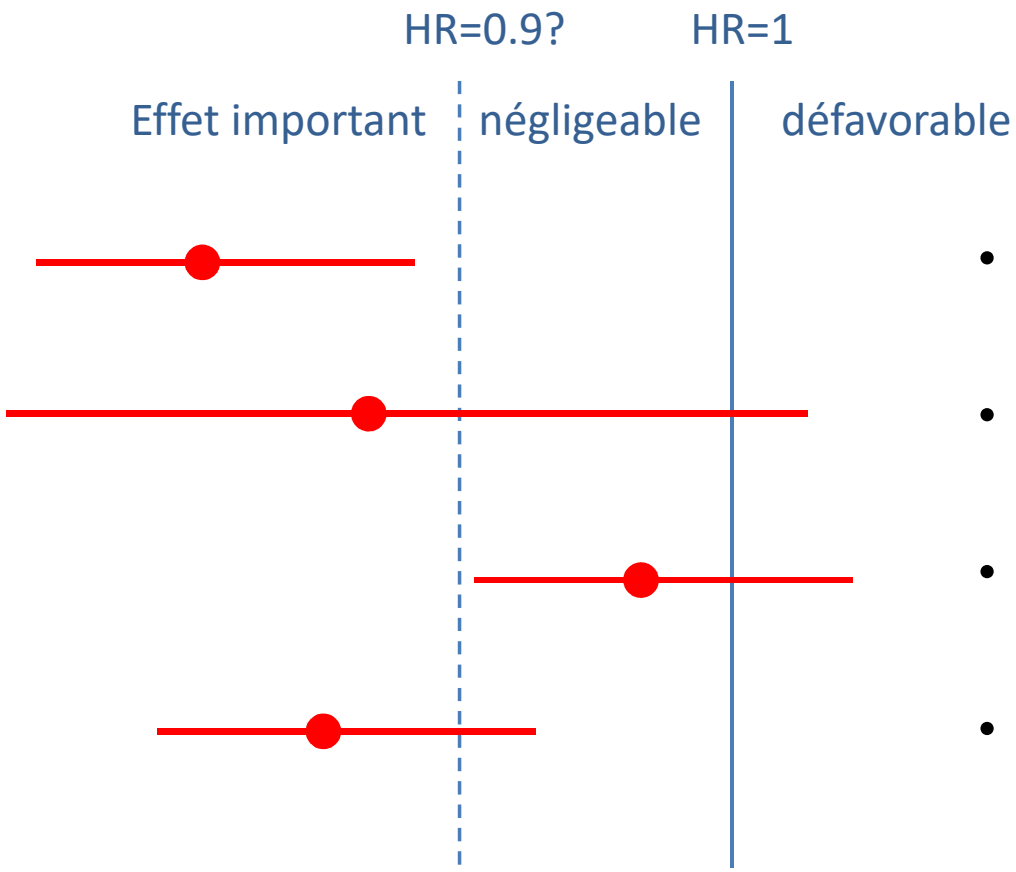
Intervalle de confiance

estimateur

Incertitude(s)

- Dans un essai clinique d'un nouveau traitement on trouve un hazard ratio de mortalité de 0.67 (IC95% 0.50 – 0.90, p<0.05)
- Interprétation du « chercheur-testeur »:
  « le traitement réduit la mortalité d'un tiers, et c'est statistiquement significatif! »
- Interprétation du « chercheur-estimateur »:
  « la meilleure estimation est que le traitement réduit la mortalité d'un tiers, mais cela pourrait aussi être une réduction de moitié, ou une réduction de 10% »
- Interprétation du « chercheur-personnaliseur »:
  « la meilleure estimation est que le traitement réduit la mortalité en moyenne d'un tiers, mais cela pourrait être une réduction de moitié, ou de 10%. En plus on ne sait pas si l'effet du traitement sur vous, cher patient, correspond à la moyenne, ou si l'effet sera plus ou moins fort »

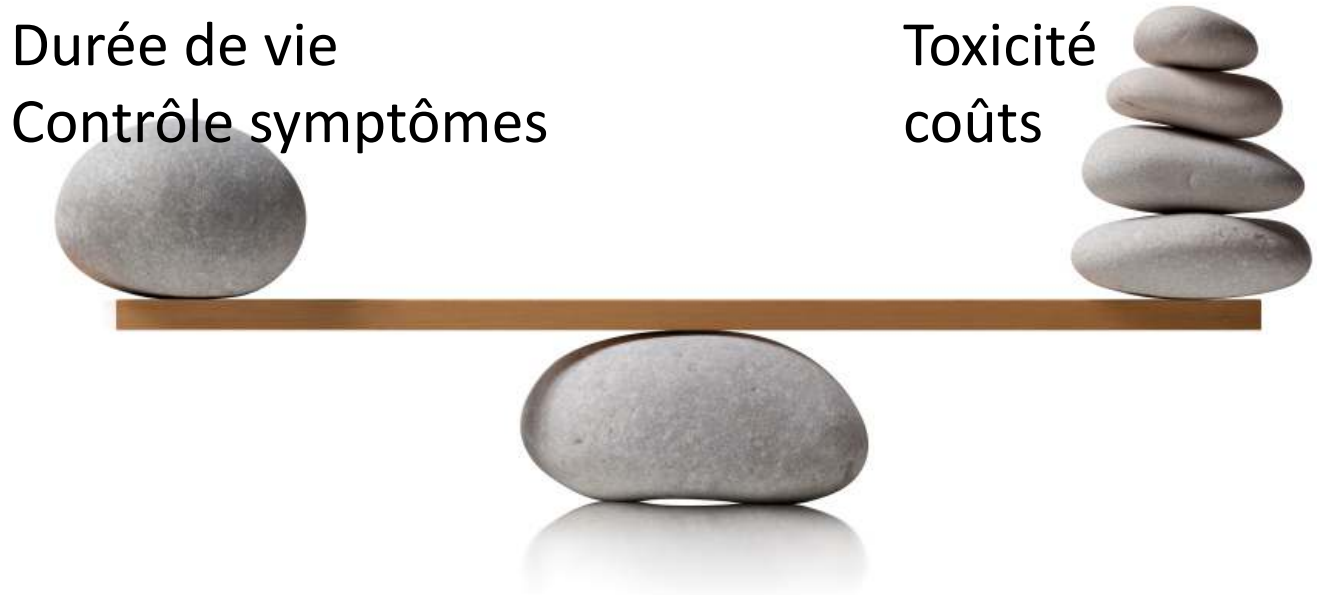
Implications des valeurs « compatibles »



- Significatif et important
- Imprécis et non-significatif
- Modeste et non-significatif
- Significatif <u>mais ambigu</u>!

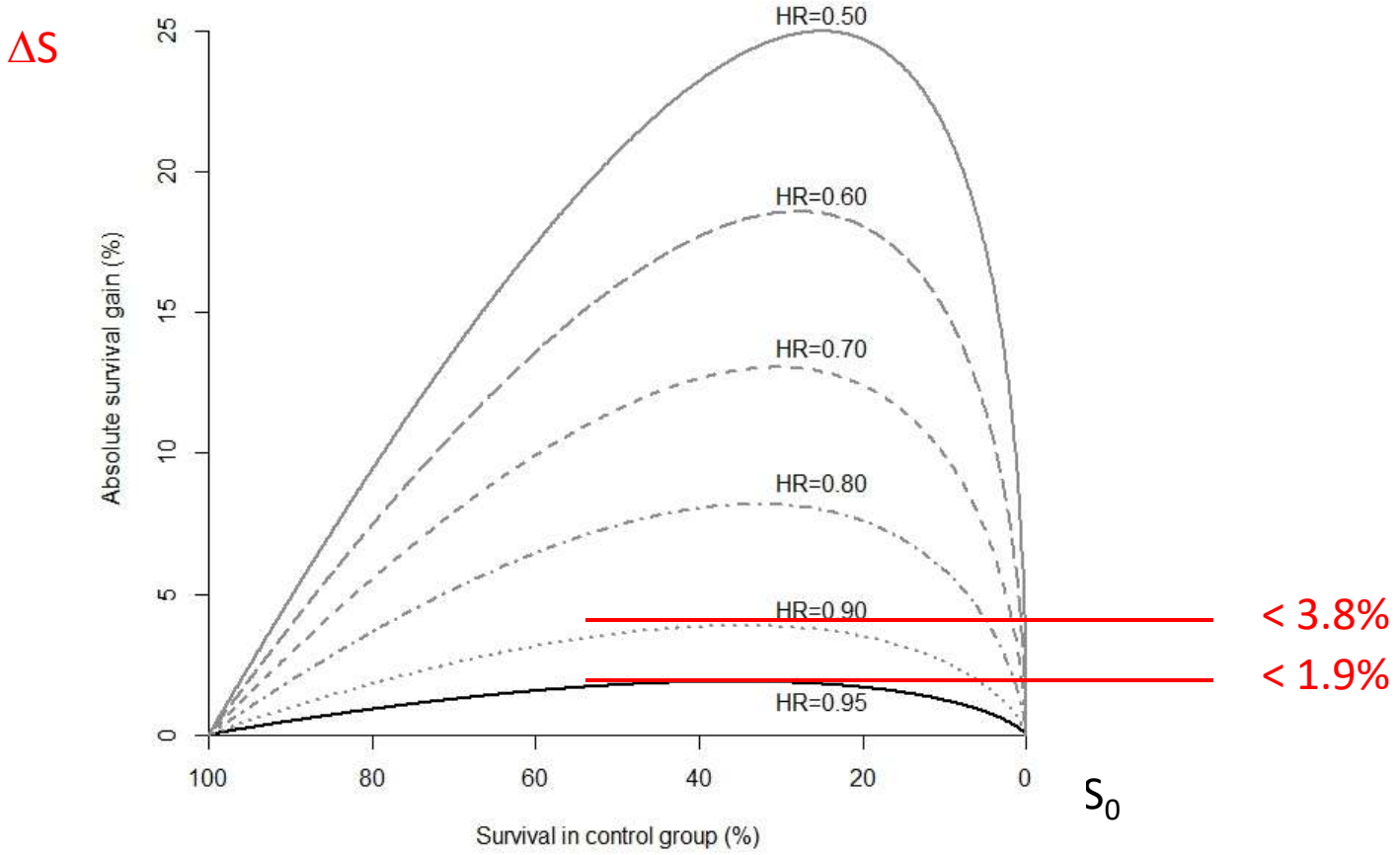
Une réduction de mortalité peut-elle être négligeable?

- Ca dépend!
- Si le traitement est bon marché et n'a pas d'effet indésirable, non
- Si le traitement est toxique, et/ou (très) cher, oui
- Le trade-off dépend des priorités et préférences des patients; il faudrait une analyse décisionnelle détaillée pour chaque cas



HR et ΔS

- La pesée des intérêts nécessite des risques absolus (et non relatifs)
- Si les hazards sont proportionnels: $HR = \ln(S_1)/\ln(S_0) = \ln(S_0+\Delta S)/\ln(S_0)$



Etude des publications

T Perneger, P Brindel, C Combescure, A Gayet-Ageron

- Arrive-t-il souvent que les études qui montrent qu'un traitement améliore significativement la survie (hazard ratio <1, p<0.05) soient aussi compatibles avec des bénéfices cliniquement négligeables?
- Définition de « cliniquement négligeable » : HR dépasse 0.90 ou 0.95

Etudes analysées

- Essais randomisés publiés dans NEJM, Lancet, Lancet Oncology, J Clinical Oncology, JAMA, JAMA Oncology entre janvier 2009 et juin 2019
- Traitements de maladies oncologiques, chimio ou radiothérapie
- Traitements démontrant un bénéfice significatif ($p<0.05$) pour la survie (overall survival)
- Exclusions: prévention, non-infériorité, supériorité du traitement standard, analyse groupée de plusieurs RCT ou méta-analyse
- Données recueillies: HR et son IC
- 1489 articles screenés
- 226 articles et 234 HR retenus (8 avaient 2 HR: RCT à 3 bras, RCT factoriels, co-primary outcomes)
- Cancers: hémato (38), poumon (31), sein (28), prostate (22), colorectal (19), estomac/oesophage (17), peau (16), pancréas (10), foie (10), ovaires/utérus (10), tête/cou (7), urinaire (7), cerveau (6), sarcome (5)

Descriptif des études

	Hazard ratios N (%)
Overall	234 (100)
Publication year:	
2009-2011	55 (23.5)
2012-2014	62 (26.5)
2015-2017	71 (30.3)
2018-2019	46 (19.7)
Journal:	
New England Journal of Medicine	76 (32.5)
Lancet	41 (17.5)
Lancet Oncology	50 (21.4)
Journal of Clinical Oncology	57 (24.4)
JAMA	5 (2.1)
JAMA Oncology	5 (2.1)

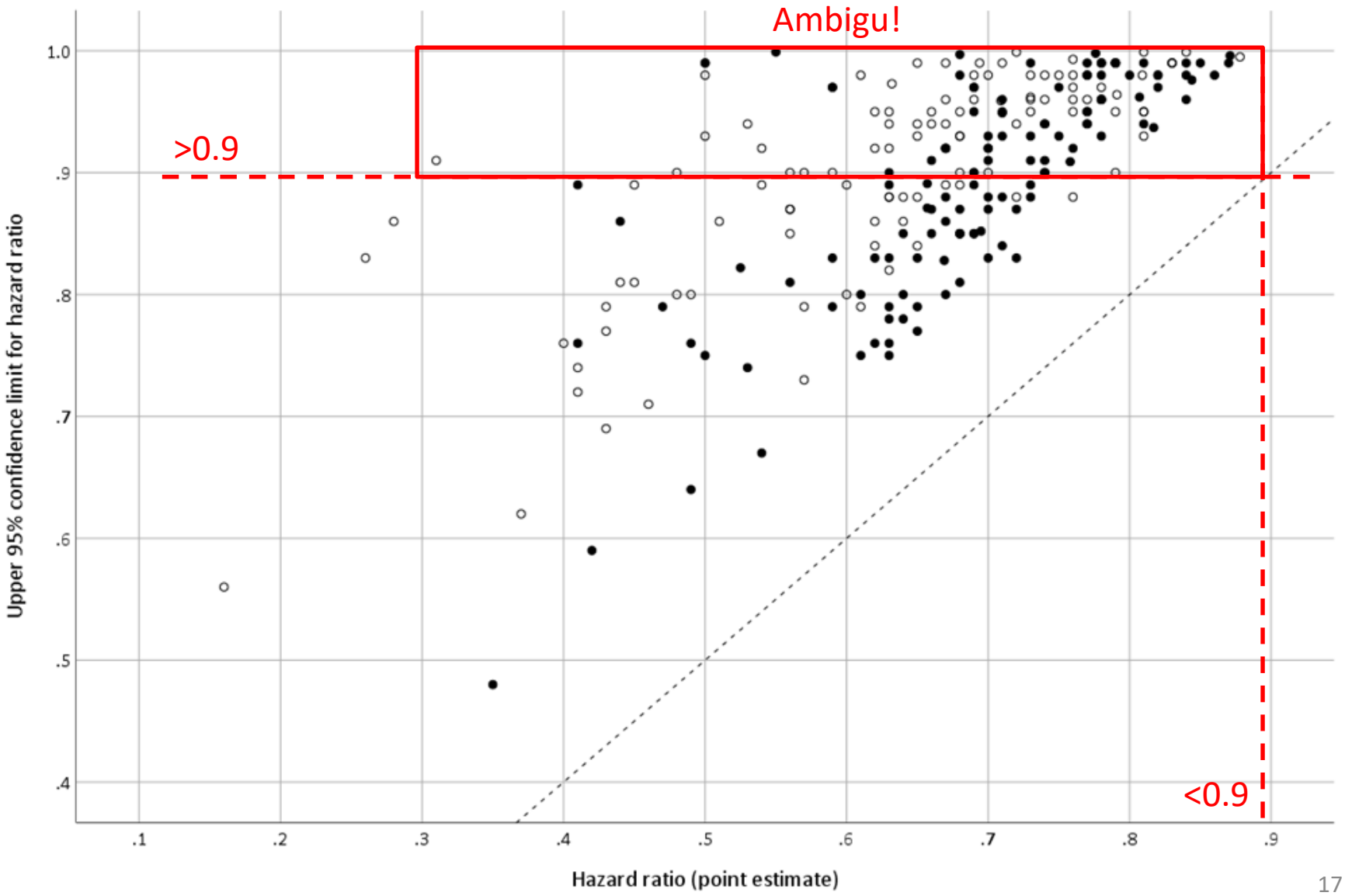
	Hazard ratios N (%)
Sample size	
65-399	78 (33.3)
400-799	90 (38.5)
800-3105	66 (28.2)
Drug development phase:	
Phase II	20 (8.5)
Phase III	209 (89.3)
Unclassified	5 (2.1)
Overall survival outcome:	
Primary (or co-primary)	117 (50.0)
Secondary	117 (50.0)

Distributions des HR

	Hazard ratio (point estimate)	Lower 95% confidence limit	Upper 95% confidence limit
Mean (SD)	0.664 (0.124)	0.500 (0.143)	0.897 (0.089)
Quartiles	0.62, 0.68, 0.76	0.44, 0.52, 0.60	0.85, 0.92, 0.97
Range	0.16 to 0.88	0.05 to 0.77	0.48 to 0.99
Categories, N (%)			
Up to 0.199	1 (0.4)	5 (2.1)	0
0.20 to 0.299	2 (0.9)	21 (9.0)	0
0.30 to 0.399	3 (1.3)	22 (9.4)	0
0.40 to 0.499	20 (8.5)	49 (20.9)	1 (0.4)
0.50 to 0.599	26 (11.1)	73 (31.2)	2 (0.9)
0.60 to 0.699	81 (34.6)	50 (21.4)	4 (1.7)
0.70 to 0.799	75 (32.1)	14 (6.0)	24 (10.3)
0.80 to 0.899	26 (11.1)	0	65 (27.8)
0.90 to 0.949	0	0	50 (21.4)
0.95 to 1.00	0	0	88 (37.6)

	Hazard ratios N (%)	Upper HR 95% confidence limit	
		≥0.90	≥0.95
Overall	234 (100)	59.0%	37.6%
Publication year:		(p=0.82)	(p=0.95)
2009-2011	55 (23.5)	60.0%	40.0%
2012-2014	62 (26.5)	62.9%	37.1%
2015-2017	71 (30.3)	54.9%	35.2%
2018-2019	46 (19.7)	58.7%	39.1%
Journal:		(p=0.013)	(p=0.019)
New England Journal of Medicine	76 (32.5)	43.4%	22.4%
Lancet	41 (17.5)	61.0%	43.9%
Lancet Oncology	50 (21.4)	66.0%	38.0%
Journal of Clinical Oncology	57 (24.4)	73.7%	50.9%
JAMA	5 (2.1)	40.0%	40.0%
JAMA Oncology	5 (2.1)	60.0%	60.0%
Sample size		(p=0.40)	(p=0.068)
65-399	78 (33.3)	55.1%	28.2%
400-799	90 (38.5)	64.4%	45.6%
800-3105	66 (28.2)	56.1%	37.9%
Drug development phase:		(p=0.27)	(p=0.76)
Phase II	20 (8.5)	45.0%	38.3%
Phase III	209 (89.3)	60.8%	30.0%
Unclassified	5 (2.1)	40.0%	40.0%
Overall survival outcome:		(p=0.017)	(p=0.015)
Primary (or co-primary)	117 (50.0)	51.3%	29.9%
Secondary	117 (50.0)	66.7%	45.3%

Limite supérieure vs HR



Distribution des ΔS quand S_0=0.5

Gains en %	Absolute survival gain (point estimate)	Lower 95% confidence limit	Upper 95% confidence limit
Mean (SD)	13.4 (5.7)	3.8 (3.5)	21.0 (7.3)
Quartiles	9.0, 12.4, 15.2	1.0, 2.8, 5.5	15.8, 19.7, 23.8
Range	4.4 to 39.5	0 to 21.7	8.5 to 46.6
Categories, N (%)			
Less than 1%	0	51 (21.8)	0
1% to 1.99%	0	38 (16.2)	0
2% to 3.99%	0	59 (25.2)	0
4% to 7.99%	33 (14.1)	62 (26.5)	0
8% to 15.99%	149 (63.7)	21 (9.0)	64 (27.4)
16% to 31.99%	49 (20.9)	3 (1.3)	146 (62.4)
32% and more	3 (1.3)	0	24 (10.3)

En parle-t-on?

- 20 mentions d'incertitude (sur 226): cross-over, survival « not mature », etc
- Trois mentions de l'intervalle de confiance
  - Given the small size of the trial and the large CIs around HRs observed, a larger trial would be needed to give more accurate estimates of the true benefit
  - The phase II nature of the trial, however, requires caution in interpretation of the results, because the probability of unstable estimates of treatment effect and false-positive results increases with small sample size
  - We note also that for overall survival, the number of events is small, and confidence intervals are wide
- Aucune mention de la limite supérieure de l'IC

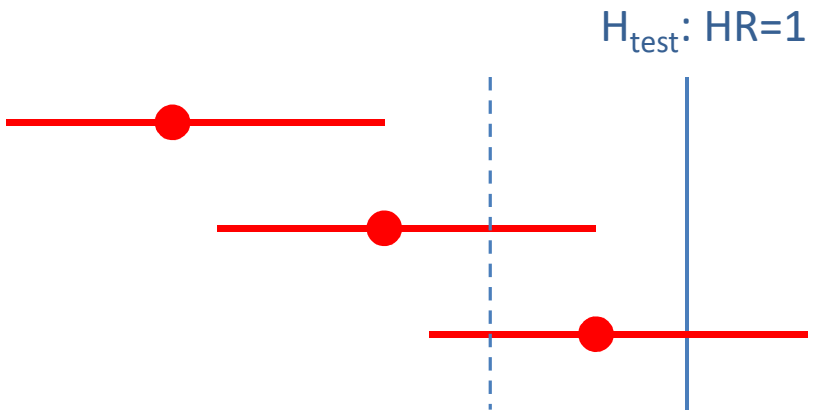
Bilan des RCT en oncologie

- Tous les HR (point estimates) du bénéfice de survie sont cliniquement importants, <0.9
- Une grande partie des limites supérieures du HR sont compatibles avec un bénéfice trivial:
  - Médiane 0.92
  - 59.0% >0.90 (51.3% des primary outcomes)
  - 37.6% >0.95 (29.9% des primary outcomes)
- Tous les ΔS (point estimates) sont cliniquement importants, $\geq$4%
- Une grande partie des limites inférieures de ΔS sont compatibles avec un bénéfice trivial:
  - Médiane 2.8%
  - 38.0% <2%
  - 21.8% <1%

Questions

- Cliniciens: Comment communiquer aux patients les résultats des études et l'incertitude concernant l'effet des traitements?
- Editeurs: quelles exigences pour les articles rapportant des RCT? Ajouter des critères à CONSORT?
- Chercheurs: comment planifier les RCT pour réduire l'incertitude concernant l'effet du traitement?

Supériorité ou « super-supériorité » ?



- Significatif
- Significatif <u>mais ambigu</u>
- Non-significatif

- Significatif
- Non-significatif
- Non-significatif

Farrington CP, Manning G. Test statistics and sample size formulae for comparative binomial trials with null hypothesis of non-zero risk difference or non-unity relative risk. Stat Med 1990;9:1447-54

Conclusion

- Bcp de traitements ayant un effet significatif sur la mortalité ont peut-être en réalité un effet négligeable (en tous cas en oncologie)
- Ces incertitudes passent (souvent) sous silence
- Que faire?