Long term toxicity of a Roundup herbicide and a Roundup-tolerant genetically modified maize

G.E. Séralini *et al.*

Food and Chemical Toxicology, 2012

Analyse statistique

Comité Scientifique du Haut Conseil des Biotechnologies

22 octobre 2012

1. Le protocole expérimental

Un échantillon de 200 rats, constitué de 100 mâles et 100 femelles, a été randomisé en 20 groupes de 10 rats de même sexe, chaque groupe recevant le même régime alimentaire.

Pour chaque sexe, un seul groupe témoin de 10 rats a été constitué. C'est donc uniquement ce même groupe de 10 rats qui est systématiquement utilisé pour être comparé aux 9 groupes expérimentaux de même sexe.

Il en résulte un manque de puissance tel qu'il est dès lors très difficile d'établir si des différences observées lors de chacune de ces 9 comparaisons sont dues à un effet du régime, ou bien simplement à une variabilité naturelle du groupe contrôle. Ainsi, si un paramètre est particulièrement élevé chez le groupe contrôle, toutes les différences entre groupes expérimentaux et contrôle iront dans le même sens et présenteront toutes une diminution du caractère considéré, sans que l'on puisse pour autant en conclure à un effet du régime. A titre d'exemple, les auteurs concluent §3.3 sans aucune précaution que « *Creatinine or clairance decreased in urine for all treatment groups in comparison to female controls (Table 3)*» alors qu'il est très probable que ce sont des valeurs particulièrement élevées de ces paramètres dans le groupe contrôle qui expliquent ces différences.

Aucun calcul préalable du nombre de sujets nécessaires pour mettre en évidence un effet jugé biologiquement significatif n'est mentionné. Un tel calcul aurait été particulièrement pertinent pour évaluer la quantité d'information, par exemple sur les durées de vie ou les nombres de tumeurs, que l'on peut

espérer obtenir avec le protocole mis en place. Il aurait ainsi été possible d'évaluer quantitativement le manque d'intérêt d'utiliser une souche de rats qui développe naturellement des tumeurs avec une grande probabilité. En effet, plus le risque de développer naturellement des tumeurs est élevé, et plus il faudra d'animaux dans les groupes expérimentaux pour mettre en évidence une augmentation significative du nombre de tumeurs dû au régime. Considérons à titre d'exemple 2 souches de rats A et B pour qui le risque de développer naturellement des tumeurs au cours d'une période de temps donnée est de 10% pour A et 60% pour B. Les tableaux ci-dessous indiquent le nombre de rats nécessaire dans un groupe expérimental pour mettre en évidence une augmentation du nombre de tumeurs de 10%, 20% ou 30%, lorsque le risque d'erreur de première espèce α et celui de seconde espèce β sont tous les deux égaux à 5% ou 10%. On voit ainsi que l'utilisation de la souche B nécessite plus d'animaux que la souche A, quels que soient les risques d'erreurs donnés et lorsque l'augmentation à mettre en évidence est inférieure à 30%.

Souche de rats

Effet à mettre en évidence	**A** p=10%	**B** p=60%
+ 10%	135	248
+ 20%	41	60
+ 30%	24	24

$\alpha=\beta=5\%$

Souche de rats

Effet à mettre en évidence	**A** p=10%	**B** p=60%
+ 10%	86	156
+ 20%	25	36
+ 30%	15	15

$\alpha=\beta=10\%$

Table 1 : nombres de rats nécessaires dans un groupe expérimental pour mettre en évidence une augmentation donnée du nombre de tumeurs dans deux souches A et B (le risque de développer naturellement des tumeurs au cours d'une période de temps donnée est de 10% pour A et 60% pour B) et pour différents risque d'erreur de première et seconde espèces[1].

Plus généralement, le plan d'analyse statistique n'est pas mentionné. Les auteurs de l'étude semblent avoir effectué leur étude statistique en fonction des résultats obtenus, ce qui est en totale contradiction avec les règles élémentaires de bonnes pratiques statistiques. En effet, le degré de signification d'une différence observée pour un paramètre donné n'est pas du tout le même, suivant que ce paramètre a été sélectionné a priori (avant d'obtenir les résultats), ou bien a posteriori (parmi les paramètres présentant le plus de différences). Les auteurs présentent ainsi Figure 5-B les 4 paramètres biochimiques et les 2 hormones qui présentent le plus de différences, parmi le groupe qui présente le plus de différences. Ce choix a été fait a posteriori, *i.e.* une fois les résultats obtenus : il est donc attendu que certaines comparaisons parmi les 18x48=864 présentent des différences qui peuvent sembler importantes. La présentation brute de ces

[1] Le risque de première espèce α est le risque de conclure à une augmentation alors qu'il n'y a pas de différence ; le risque de seconde espèce β consiste à ne pas détecter d'augmentation alors qu'elle existe.

résultats partiels peut alors être trompeuse pour un lecteur non spécialiste des comparaisons multiples, qui risque de considérer à tort ces différences observées comme représentatives des différences entre groupes expérimentaux et groupes contrôle.

2. Analyse descriptive des résultats

Le corps de l'article se limite essentiellement à une description des résultats obtenus lors de cette étude chez les différents groupes de 10 rats (courbes de mortalité, pathologies anatomiques,…).

Les commentaires des auteurs illustrent de façon partielle ce qui a été observé. On peut ainsi lire §3.1 "*Before this period, 30% control males (three in total) and 20% females (only two) died spontaneously, while up to 50% males and 70% females died in some groups on diets containing the GM maize (Fig. 1).*". Mais si quelques groupes expérimentaux de mâles présentent en effet un taux de mortalité de 50% (5 rats morts) à la date de 600 jours, les groupes de mâles ayant été nourris avec des doses supérieures de maïs GM ou de Roundup présentent des taux de mortalités de seulement 10% (1 rat mort). Cette différence observée n'est pas mentionnée. D'autre part, choisir d'extraire un taux de mortalité à la date particulière de 600 jours est totalement arbitraire : le taux de mortalité de 30% chez les mâles du groupe témoin à 600 jours passe à 50% vers 620 jours.

On comprend mal la valeur statistique à accorder aux différentes photos (rats, organes, tumeurs,…) puisque certains rats seulement sont représentés. On se demande ainsi si les rats sélectionnés sont représentatifs de leur groupe. Les groupes témoins devraient alors également être représentés.

3. Inférence statistique

C'est la statistique inférentielle qui permet d'évaluer les incertitudes et les probabilités de se tromper en concluant à la présence ou à l'absence d'effets. C'est-à-dire si les différences observées peuvent être expliquées par un effet du régime, ou bien simplement par les fluctuations aléatoires d'échantillonage. En d'autres termes, la question qui se pose naturellement dans ce genre d'étude concerne la reproductibilité des résultats observés : si l'on répète la même expérience dans les mêmes conditions, quelles seront les chances d'obtenir des résultats similaires à ceux observés ici ?

En ce qui concerne les études de durées de vie et de nombres de tumeurs, les auteurs ont totalement négligé cet aspect de la statistique, tout en s'autorisant à des interprétations non justifiées de leurs résultats expérimentaux. On lit ainsi p 8-9 :

- ***All treatments in both sexes enhanced large tumor incidence*** *by 2–3-fold in comparison to our controls but also for the number of mammary tumors…*
- *Suffering inducing euthanasia and deaths corresponded mostly in females to the development of large mammary tumors.* ***These appeared to be clearly related to the various treatments when compared to the control groups.***"

On ne trouve pas un seul argument statistique dans cet article qui démontre l'existence de telles relations de cause à effet. On ne trouve pas la moindre analyse statistique qui mette en évidence une différence statistiquement significative des durées de vie et des nombres de tumeurs entre groupe expérimentaux et groupes contrôle.

Concernant les paramètres biochimiques, on peut lire dans la conclusion de l'article :

- *The results of the study presented here* ***clearly demonstrate*** *that lower levels of complete agricultural glyphosate herbicide formulations, at concentrations well below officially set safety limits,* ***induce severe hormone-dependent mammary, hepatic and kidney disturbances****.*
- *Altogether, the* ***significant*** *biochemical disturbances and physiological failures documented in this work* ***confirm the pathological effects*** *of these GMO and R treatments in both sexes, with different amplitudes.* [2]

De telles affirmations méritent d'être rigoureusement justifiées et validées. Or, il est ici absolument impossible de conclure de façon définitive à la toxicité du NK603 sur la base de données aussi limitées.

4. Analyse de survie

4.1 Durée de vie

Les auteurs expliquent §3.1 que *"Control male animals survived on average 624 ± 21 days, whilst females lived for 701 ± 20"*. Ces valeurs sont le résultat d'un calcul incorrect puisque nous sommes en présence de données censurées (on ignore quand les animaux encore vivants en fin d'étude seraient morts naturellement puisqu'ils ont été euthanasiés). Ce sont les moyennes et écart-types empiriques des valeurs observées non censurées et des temps de censures qui ont été calculées (comme si un rat encore vivant à T=720j était considéré comme mort à T=720j). Les résultats présentés sont donc inexacts puisque cette procédure introduit un biais en sous estimant bien sûr le temps moyen de mort, mais surtout l'erreur standard de l'estimateur. Avoir choisi de griser tout ce qui se passe après 624-21=603 jours n'est donc pas justifié puisque cette valeur est issue d'un calcul incorrect.

Un calcul correct de la distribution des durées de vie dans les différents groupes nécessite d'utiliser un modèle paramétrique, mais l'intérêt d'une telle approche est limitée avec aussi peu de données par groupe. A titre d'exemple, si l'on ajuste un modèle gaussien pour le temps de survie des mâles, les moyennes et écart-type estimées sont respectivement 626 jours et 68 jours. Pour les femelles, les moyennes et écart-type estimées sont respectivement 892 jours et 206 jours ! On ne peut, comme le font les auteurs, calculer une erreur standard pour la moyenne en divisant simplement l'écart-type par √10, comme on le ferait pour des

[2] - Les résultats de l'étude présentée ici démontrent clairement que de faibles niveaux de préparations complètes d'herbicide agricole à base de glyphosate, à des concentrations bien inférieures aux limites de sécurité fixées officiellement, induisent des troubles hormono-dépendant sévères d'ordre mammaire, hépatique et rénal.
- Au final, les perturbations biochimiques significatives et les troubles physiologiques présentés dans ce travail confirment les effets pathologiques de ces traitements OGM et R pour les deux sexes, avec des amplitudes différentes.

variables gaussiennes non censurées. A cause du phénomène de censure, la distribution de l'estimateur de la moyenne est beaucoup plus dispersée et très asymétrique, ce qui rend de plus l'utilisation de l'erreur standard peu pertinente pour calculer un intervalle de confiance.

4.2 Comparaisons entre groupes expérimentaux et groupes contrôles

L'analyse de survie réalisée dans cette étude se limite à une représentation graphique des courbes de mortalité dans chaque groupe (nombre de rats morts en fonction du temps).

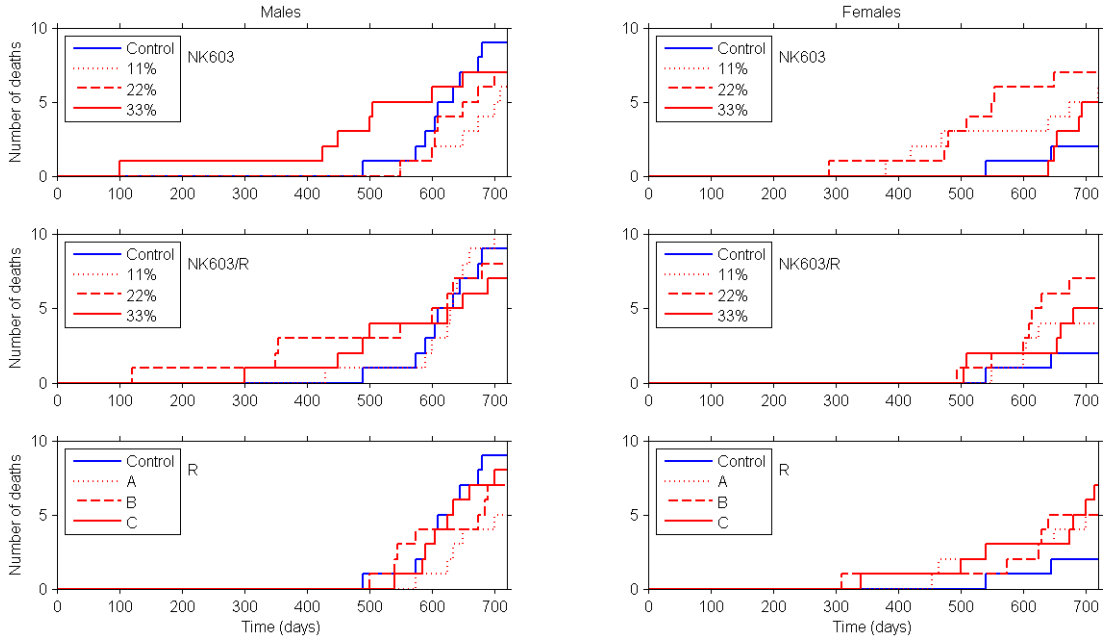

Fig. 1 Courbes de mortalité pour les 18 groupes expérimentaux et les 2 groupes contrôles.

Un simple examen des courbes de mortalité observées ne permet pas de conclure à une quelconque différence au niveau de la population. Il faut pour cela mettre en œuvre une procédure statistique rigoureuse qui prend en compte la variabilité des individus et donc des courbes de survie.

De nombreuses techniques statistiques existent pour comparer des courbes de survie. On peut dans un premier temps considérer les 18 comparaisons entre groupes expérimentaux et groupes témoins. Le test de rang (ou test de Wilcoxon) est un test non paramétrique qui permet de comparer les statistiques de rang de 2 échantillons.

On peut ainsi tester si les animaux d'un groupe expérimental donné ont tendance à mourir plus tôt que ceux du groupe contrôle. Par exemple :

H_0 « le régime NK603 à 11% n'a pas d'effet sur la durée de vie des rats femelles »

vs

H_1 « le régime NK603 à 11% entraîne une diminution de la durée de vie des rats femelles »

La statistique de test est définie comme la somme des rangs du groupe contrôle. Pour chacune des 18 comparaisons, cette statistique de test peut être calculée et comparée à un intervalle de prévision obtenu sous l'hypothèse nulle. Un degré de signification peut être calculé pour chacun des 18 tests effectués.

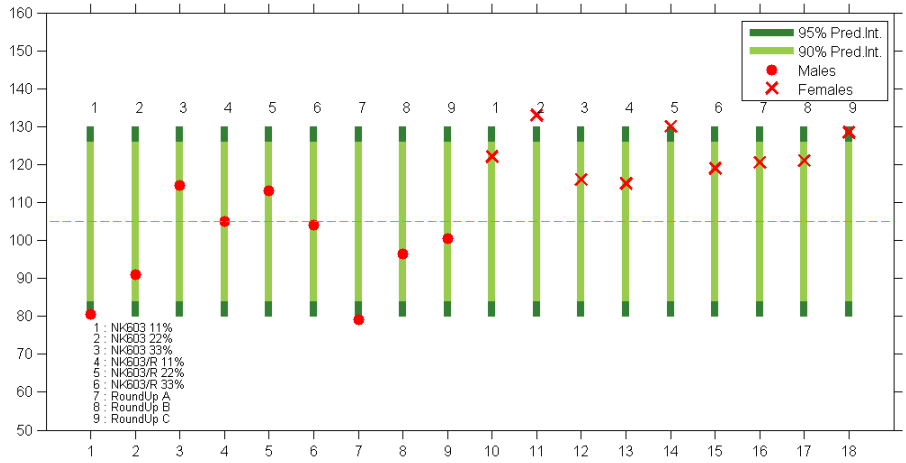

Fig. 2a Intervalles de prévision de niveau 90% et 95% des 18 statistiques de tests et valeurs observées de ces statistiques de tests.

Néanmoins, ces intervalles de prévision ne prennent pas en compte la multiplicité des comparaisons effectuées. Plutôt que mettre en œuvre un test trop conservateur (qui tendrait à conserver trop systématiquement l'hypothèse nulle) on peut estimer par simulation, ou par permutation, la distribution de probabilité des 18 statistiques utilisées pour ce test. La figure ci-dessous montre les intervalles de prévision des 18 statistiques de tests lorsqu'elles sont classées par ordre décroissant (ces intervalles sont estimés par simulation). Les 18 statistiques de tests sont toutes à l'intérieur des intervalles de prévision de niveau 90% correspondants :

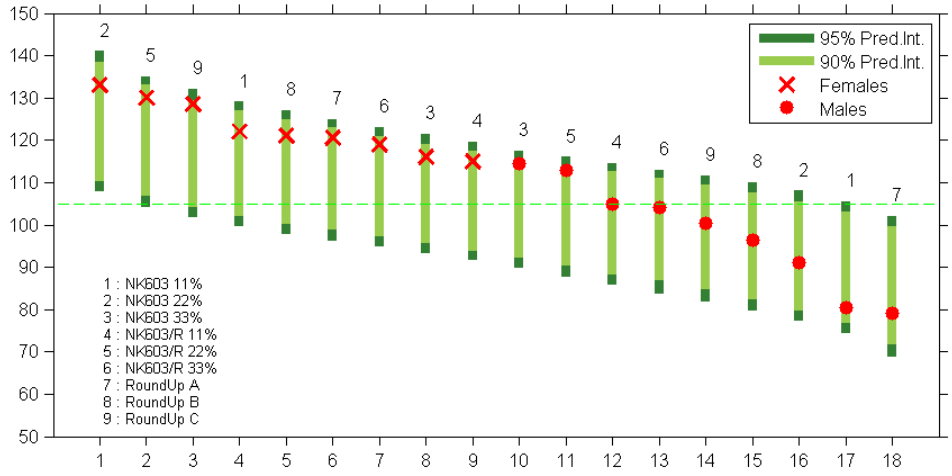

Fig. 2b Intervalles de prévision de niveau 90% et 95% des 18 statistiques de tests ordonnées par ordre décroissant et valeurs observées de ces statistiques de tests.

On peut de plus estimer le degré de signification de chaque comparaison comme un quantile empirique. Le tableau présente pour chacun des 18 tests effectués la statistique de test, *i.e.* la somme des rangs du groupe contrôle (sous l'hypothèse nulle, la valeur attendue de cette statistique est (1+2+…+19+20)/2=105), le degré de signification du test de rang correspondant, le degré de signification corrigé qui prend en compte la multiplicité des tests effectués (par simulation ou par permutation). Les groupes qui ont tendance à mourir avant le groupe contrôle (statistique de test supérieure à 105) sont en rouge et les autres en bleu :

Groupe expérimental	Statistique de test	degré de signification	degré de signification corrigé (simulation)	degré de signification corrigé (permutation)
M - NK603 11%	80.5	0.972	0.879	0.895
M - NK603 22%	91	0.865	0.665	0.624
M - NK603 33%	114.5	0.247	0.151	0.073
M - NK603/R 11%	105	0.515	0.385	0.291
M - NK603/R 22%	113	0.285	0.167	0.072
M - NK603/R 33%	104	0.545	0.368	0.263
M - RoundUp A	79	0.978	0.819	0.808
M - RoundUp B	96.5	0.753	0.514	0.452
M - RoundUp C	100.5	0.648	0.422	0.343
F - NK603 11%	122	0.072	0.199	0.174
F - NK603 22%	133	**0.011***	0.158	0.166
F - NK603 33%	116	0.176	0.195	0.124
F - NK603/R 11%	115	0.188	0.185	0.104
F - NK603/R 22%	130	**0.021***	0.122	0.104
F - NK603/R 33%	119	0.116	0.152	0.092
F - RoundUp A	120.5	0.092	0.140	0.098
F - RoundUp B	121	0.085	0.178	0.139
F - RoundUp C	128.5	**0.029***	0.087	0.067

Table 2 Les 18 statistiques de tests (test de Wilcoxon) et les degrés de signification associés.

Le groupe qui présente le plus de différences en ce qui concerne la survie est le groupe des femelles nourries avec un régime contenant 22% de maïs NK603 non traité. Le degré de signification de ce test est de 1.1% (*i.e.* la probabilité d'obtenir une statistique de test supérieure ou égale à 133 sous l'hypothèse nulle est de 1.1%). En prenant compte de la multiplicité des tests, cette probabilité est de 15.8% lorsqu'elle est estimée par simulation et de 16.6% lorsqu'elle est estimée par permutation : la probabilité que la plus grande statistique de test parmi 18 soit supérieure ou égale à 133 sous l'hypothèse nulle est d'environ 16%. Ce tableau nous permet donc de conclure que :

- **Aucune différence observée entre les courbes de survie des groupes expérimentaux et celles des groupes témoins n'est statistiquement significative.**

Enfin, une étude de puissance par simulation montre, à titre d'exemple, que si la mort de 5 rats d'un groupe expérimental de 10 rats survient avant celle des rats contrôle, le degré de signification du test est de 8%. Ce degré de signification du test n'est plus que de 2% (resp. 0.8%) si ce sont 6 (resp. 7) rats qui meurent avant les témoins.

4.3 Utilisation de données de référence

Le manque de puissance, dû à un très faible effectif des groupes contrôle (10 rats), interdit bien entendu de conclure de façon formelle à la présence ou à l'absence d'un effet du régime sur la mortalité, en particulier chez les femelles. Ce manque de puissance peut être compensé par l'introduction d'informations à priori sur le comportement attendu des groupes contrôle. Ainsi, des données de mortalité de la souche SD fournies par la société Harlan sont disponibles et peuvent venir compléter l'information apportée par l'expérimentation. Bien sûr, ces études n'ont pas été réalisées exactement dans les mêmes conditions que l'étude qui nous intéresse et un biais peut donc être introduit en utilisant cette information a priori. Au contraire, l'information apportée par les groupes contrôles n'est pas biaisée si tous les groupes ont été suivis dans les mêmes conditions, mais comme nous l'avons vu, cette information est entachée d'une très grande variabilité. La combinaison de l'information « a priori » fournie par l'éleveur et de celle fournie par les données de l'expérience permet un bon compromis « biais-variance ».

Ici, les données fournies par la société Harlan indiquent un taux de survie à 2 ans de 32% pour les mâles et 48% pour les femelles. Pour chaque sexe, le nombre de rats en vie après 2 ans est une variable aléatoire binomiale dont on peut construire des intervalles de prévision de niveau 90% ou 95%[3]

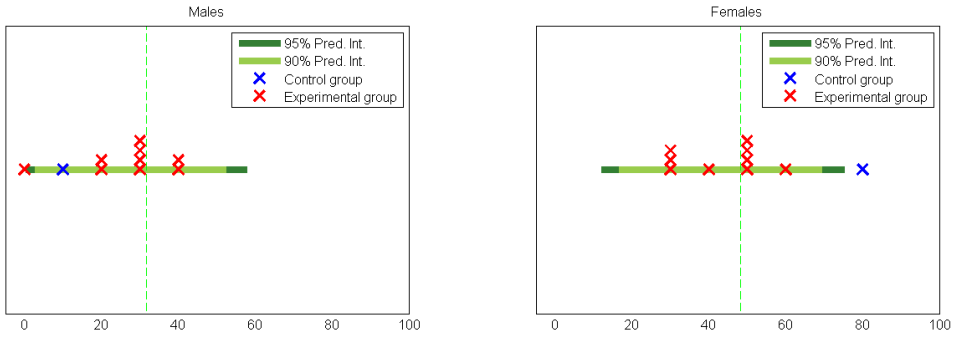

Fig.3 Intervalles de prévision de niveau 90% et 95% des taux de survie à 2 ans obtenus à partir des données Harlan et taux de survie observés dans les groupes expérimentaux et les groupes contrôle.

Les groupes expérimentaux sont très globalement distribués à l'intérieur de ces intervalles de prévision. Sur 18 groupes, il n'y a rien d'anormal à ce qu'une observation soit au bord d'un intervalle de prévision de niveau 95%. Les taux de survie observés à 2 ans sont tout à fait compatibles avec les données de référence fournies par l'éleveur, lorsque les rats sont élevés en condition normale.

- **L'utilisation de données de référence fournies par l'éleveur confirme que l'on ne peut expliquer les différences observées dans les courbes de survie entre groupes expérimentaux et groupes contrôle par un effet régime.**

[3] Pour une variable binomiale, seule les quantiles d'ordre(i/n, i=0, 2,…n) peuvent être calculés directement à partir de la distribution de probabilité. On obtient n'importe quel autre quantile par interpolation linéaire.

On peut également remarquer que les taux de survie observés dans les groupes contrôle sont au contraire assez éloignés de ce que les valeurs de références laissent présager (la proportion observée de rats femelles en vie après 2 ans est hors de l'intervalle de prévision de niveau 95%). Cela confirme la fragilité statistique de résultats obtenus à partir d'aussi faibles effectifs : on ne peut tirer de ces données aucune conclusion définitive.

4.4 Moins de groupes, mais plus de puissance

Ces résultats montrent qu'il est impossible de considérer avec une puissance suffisante les 18 comparaisons possible entre groupes expérimentaux et groupes contrôle : le protocole n'est pas du tout adapté pour un objectif aussi ambitieux.

Limiter les comparaisons en regroupant certains groupes permet de mettre en œuvre des tests plus robuste et plus puissants. On peut se limiter par exemple à tester si la mortalité chez les rats témoins est plus faible qu'au sein des groupes expérimentaux. On regroupe alors, pour chaque sexe, les groupes expérimentaux et on construit une unique courbe de survie (probabilité d'être en vie au cours du temps). Les 2 groupes expérimentaux (mâles et femelles) sont maintenant chacun formés de 90 animaux : on peut donc raisonnablement approcher les vraies fonctions de survie inconnues par les courbes de survie empiriques, obtenues à partir de ces échantillons de 90 rats. On peut alors comparer ces fonctions de survie aux courbes de survie empiriques des groupes contrôles.

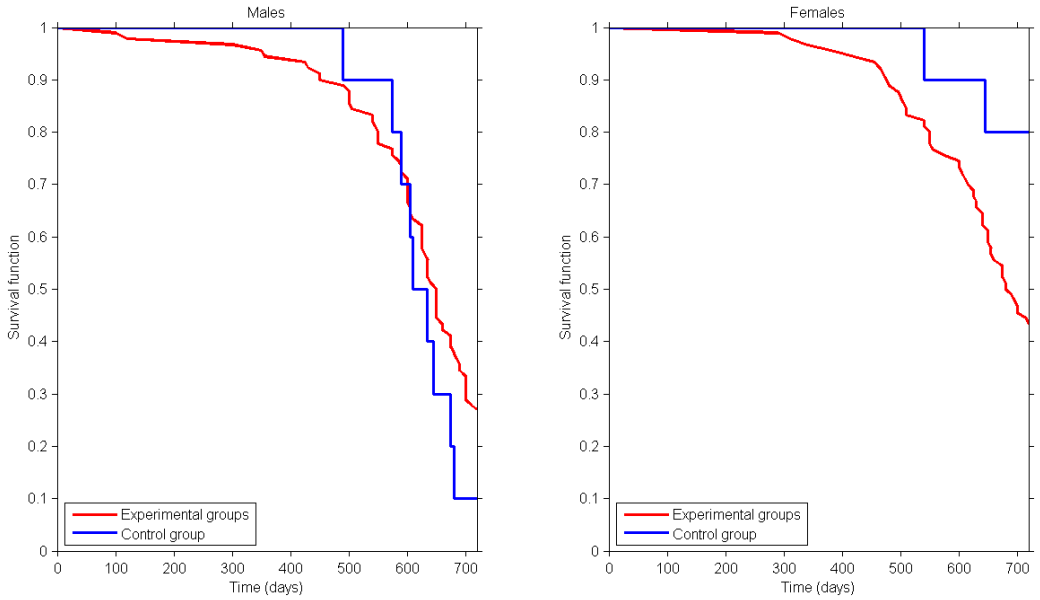

Fig. 4 Courbes de survie dans les groupes expérimentaux et groupes contrôles. Pour chaque sexe, le groupe expérimental regroupe les 9 groupes expérimentaux initiaux.

La question qui se pose est de savoir si la courbe de survie bleue (groupe contrôle observé) peut avoir été obtenue à partir de 10 rats dont la probabilité de survie est décrite par la courbe rouge.

Un intervalle de prévision des courbes de survie des 2 groupes expérimentaux (mâles et femelles) peut être facilement construit par simulation. Pour chaque sexe, on utilise la courbe de survie du groupe expérimental (rouge) pour simuler un très grand nombre (10 000 dans cet exemple) de groupes de 10 rats et leurs dates de décès. On peut alors construire les 10 000 courbes de survie empiriques obtenues à partir de ces 10 000 groupes simulés. On construit finalement un intervalle de prévision de niveau 1-α en calculant à chaque instant les quantiles empiriques d'ordre $\alpha/2$ et 1-$\alpha/2$ des 10 000 courbes de survie. Ainsi, un intervalle de prévision de niveau 90% (resp. 95%) est obtenu en calculant les quantiles empiriques d'ordre 5% (resp. 2.5%) et 95% (resp. 97.5%).

Le graphique ci-dessous représente pour chaque sexe les courbes de survie des 9 groupes expérimentaux et les intervalles de prévision de niveau 90% et 95%.

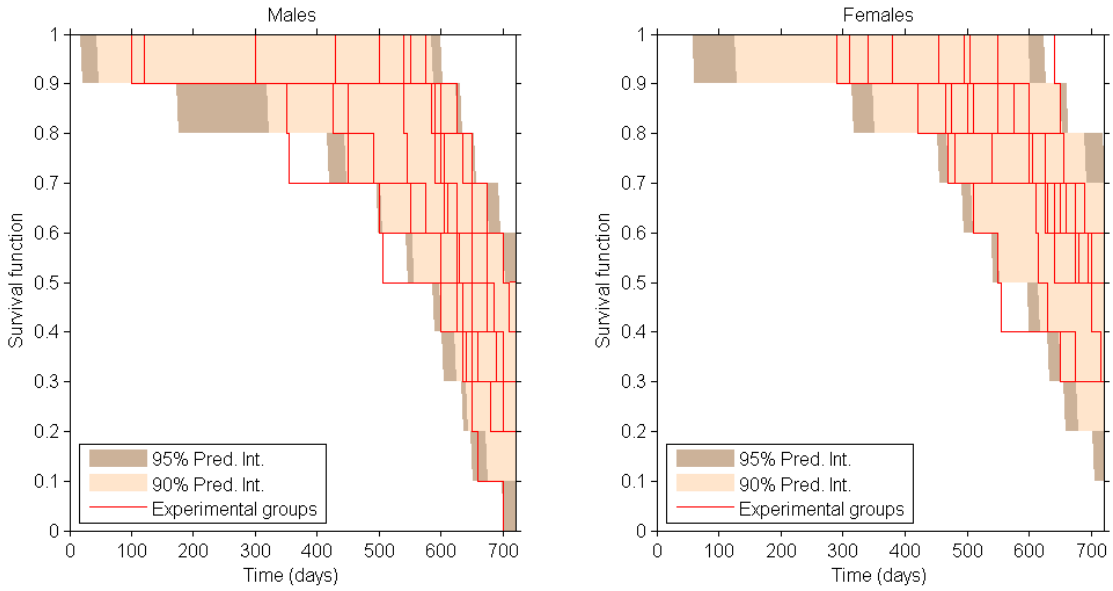

Fig. 5 Intervalles de confiance de niveau 90% et 95% de la survie du groupe expérimental et courbes de survie observées des groupes expérimentaux.

Le graphique ci-dessous représente maintenant pour chaque sexe les mêmes intervalles de prévision, mais avec cette fois-ci la courbe de survie du groupe contrôle :

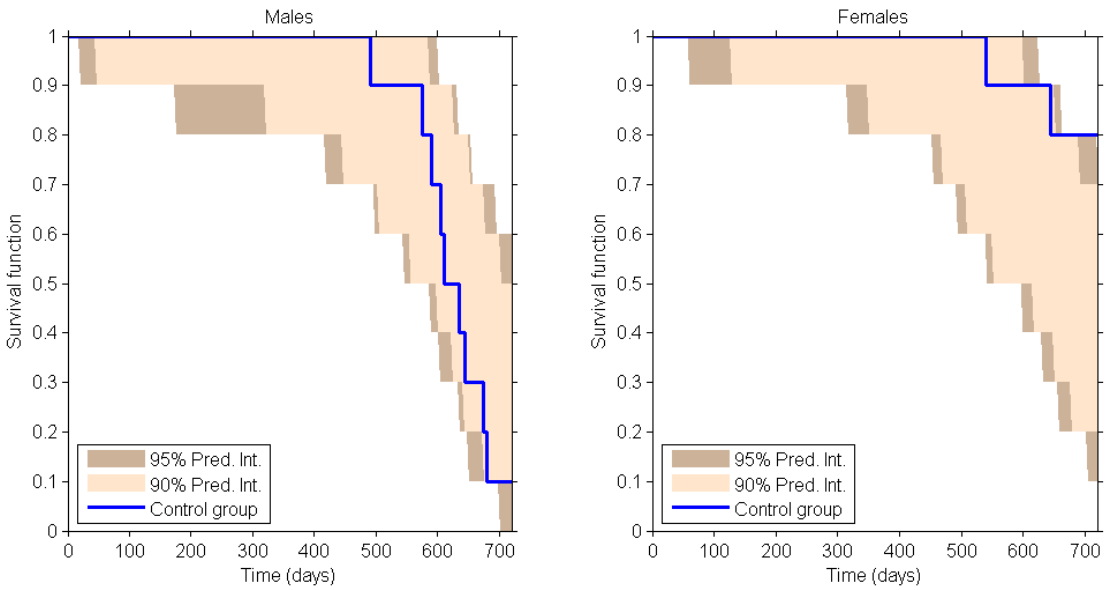

Fig. 6 Intervalles de prévision de niveau 90% et 95% de la survie du groupe expérimental et courbes de survie observées des groupes contrôle.

Les courbes de survie des groupes contrôles sont très globalement à l'intérieur de ces intervalles de prévision:

- **On ne peut donc conclure à une différence statistiquement significative entre la survie des rats contrôles et des rats tests.**

Là encore, les données de référence fournies par l'éleveur viennent étayer cette conclusion puisque les intervalles de prévision des taux de mortalité à 2 ans (pour un échantillon de taille n=90) contiennent les taux observés dans les groupes expérimentaux :

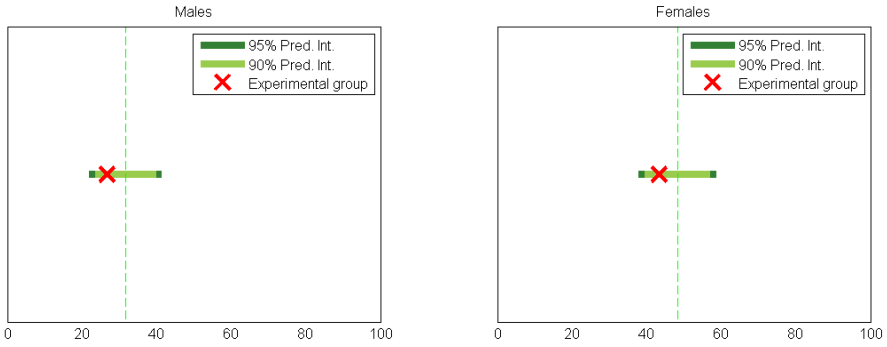

Fig. 7 Intervalles de prévision de niveau 90% et 95% des taux de mortalités à 2 ans obtenus à partir des données Harlan et taux mortalités observés dans les groupes expérimentaux (chacun de taille 90).

Le choix de regrouper les 9 groupes expérimentaux et tester si leur distribution est différente de celle du groupe contrôle est bien entendu discutable. On peut vouloir mettre en œuvre d'autres tests, et tester par exemple si la consommation d'un maïs OGM a un impact sur la survie, et ce, quel que soit la dose de maïs OGM dans le régime et quel que soit le traitement associé (RoundUp ou non). Le groupe expérimental est alors formé des 6 premiers groupes expérimentaux (NK603 et NK603/R, 11%, 22% et 33%) tandis que le groupe contrôle est formé du groupe contrôle initial (ni OGM, ni OGM/R, ni R) et des 3 groupes des rats ayant absorbé du RoundUp. Le graphique ci-dessous montre qu'il n'existe pas de différences significatives dans les survies au sein de ces groupes, aussi bien pour les mâles que pour les femelles. On ne peut donc conclure à un effet significatif du maïs NK603 sur la mortalité des rats.

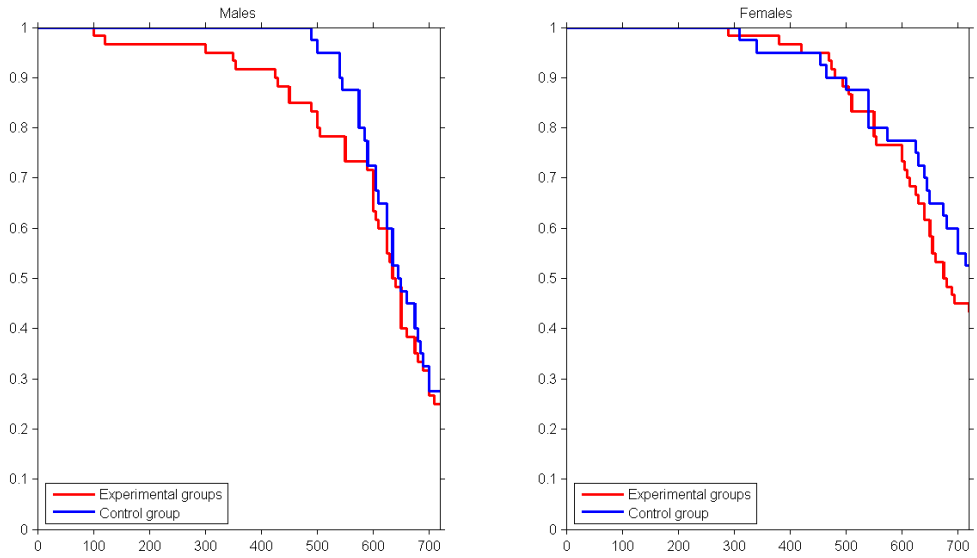

Fig. 8 Courbes de survie dans les groupes expérimentaux et groupes contrôles. Pour chaque sexe, le groupe expérimental regroupe les 6 groupes expérimentaux initiaux ayant consommé du maïs NK603 traité au RoundUp ou non traité. Le groupe contrôle contient le groupe contrôle initial et les 3 groupes ayant absorbé du RoundUp.

On peut également tester si l'absorption de RoundUp (sous forme liquide) a un impact sur la survie, et ce, quel que soit la quantité de RoundUp absorbée. Le groupe expérimental est alors formé des 3 groupes de rats ayant absorbé du RoundUp tandis que le groupe contrôle est formé du groupe contrôle initial (ni OGM, ni OGM/R, ni R) et des 6 premiers groupes expérimentaux (NK603 et NK603/R, 11%, 22% et 33%). Le graphique ci-dessous montre qu'il n'existe pas de différences significatives dans les survies au sein de ces groupes, aussi bien pour les mâles que pour les femelles. On ne peut donc conclure à un effet significatif du RoundUp sur la mortalité des rats.

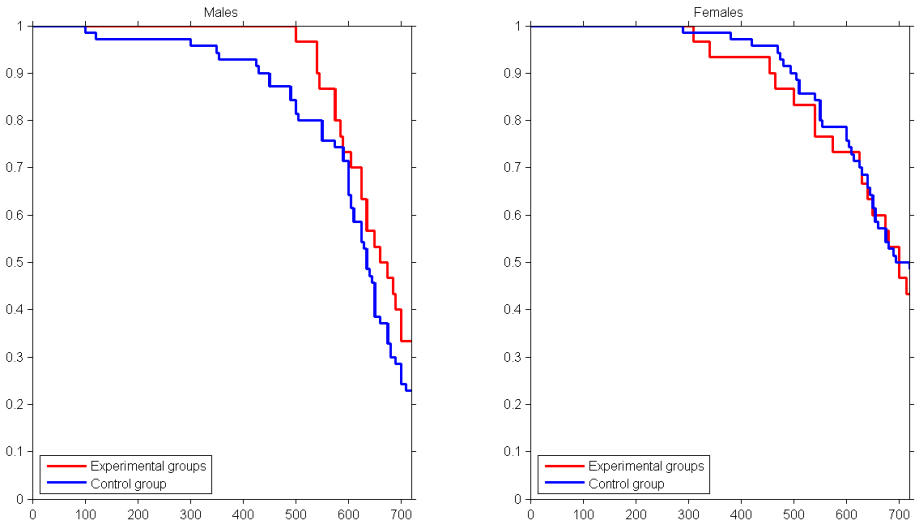

Fig. 9 Courbes de survie dans les groupes expérimentaux et groupes contrôles. Pour chaque sexe, le groupe expérimental regroupe les 3 groupes ayant absorbé du RoundUp. Le groupe contrôle contient le groupe contrôle initial et les 6 groupes expérimentaux initiaux ayant consommé du maïs NK603 traité au RoundUp ou non traité.

En conclusion, on ne peut conclure à un effet statistiquement significatif d'aucun traitement (NK603, NK603 traité au R, R) sur la survie des animaux.

5. Nombres de tumeurs

L'analyse des nombres de tumeurs réalisée dans cette étude se limite à une représentation graphique nombres de tumeurs palpables observées dans chaque groupe en fonction du temps :

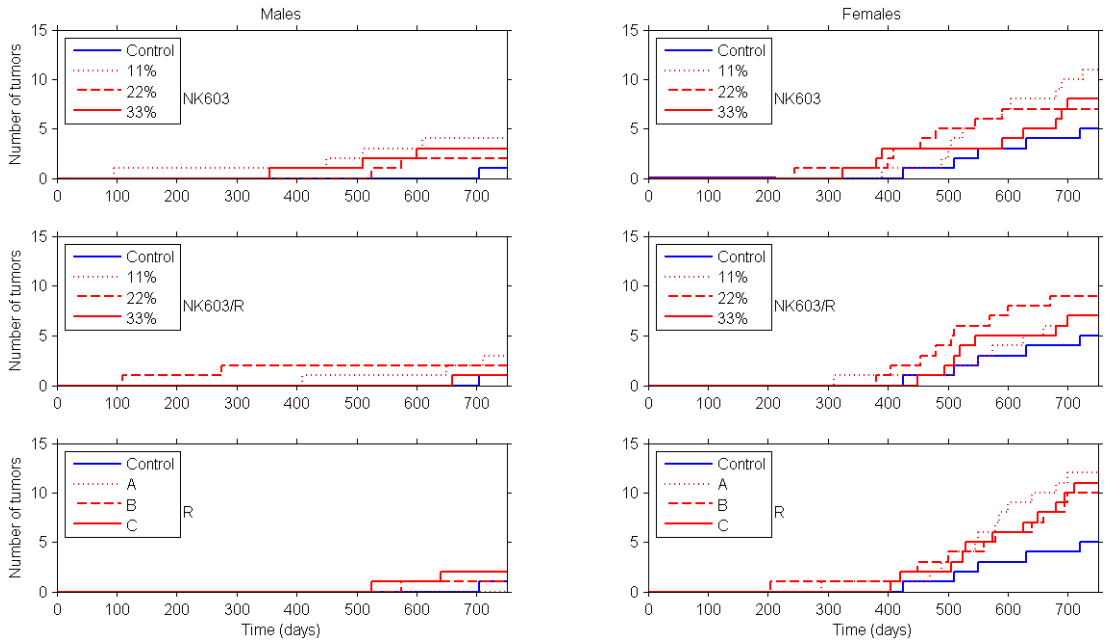

Fig. 10 Evolution des nombres de tumeurs palpables dans les 20 groupes .

Comme pour les courbes de survie, un simple examen de ces courbes observées ne permet pas de conclure à une quelconque différence au niveau de la population. Il faut pour cela mettre en œuvre un test statistique rigoureux qui prend en compte la variabilité des individus et donc des courbes de nombres de tumeurs.

Et comme précédemment, le protocole n'est pas du tout adapté pour effectuer les 18 comparaisons proposées. On peut, pour chaque sexe, regrouper les groupes expérimentaux et construire une unique courbe de nombres de tumeurs que l'on peut comparer à celle obtenue à partir du groupe contrôle.

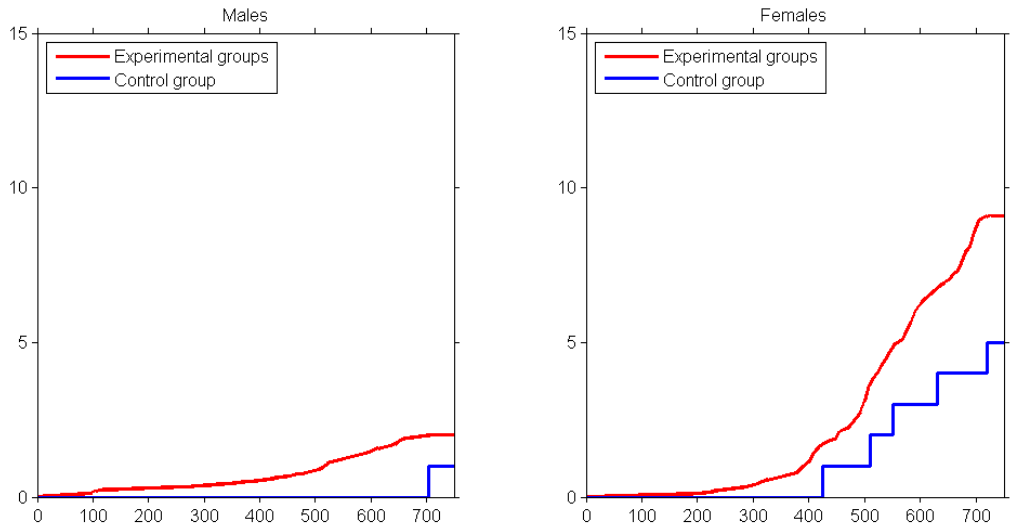

Fig. 11 Nombres de tumeurs palpables dans les groupes expérimentaux et groupes contrôles. Pour chaque sexe, le groupe expérimental regroupe les 9 groupes expérimentaux initiaux.

La question qui se pose ici est de savoir si la courbe de nombres de tumeurs bleue (groupe contrôle observé) peut avoir été obtenue à partir de 10 rats dont l'évolution du nombre de tumeurs est décrite par la courbe rouge.

Un intervalle de prévision des courbes de nombres de tumeurs des 2 groupes expérimentaux (mâles et femelles) peut être facilement construit en supposant que, pour chaque sexe, le nombre de tumeurs est un processus de Poisson non homogène dont l'intensité est donnée à chaque instant par la courbe rouge. Ainsi, un intervalle de prévision de niveau 90% (resp. 95%) est obtenu en calculant à chaque instant les quantiles d'ordre 5% (resp. 2.5%) et 95% (resp. 97.5%) d'une variable de Poisson.

Le graphique ci-dessous représente pour chaque sexe les nombres de tumeurs des 9 groupes expérimentaux et les intervalles de prévision de niveau 90% et 95% :

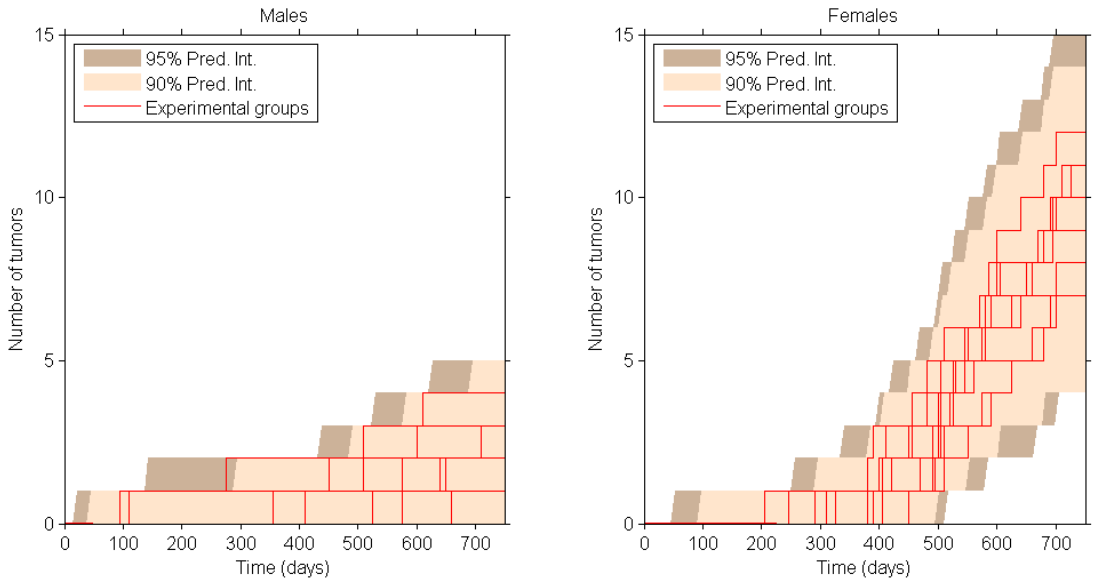

Fig. 12 Intervalles de prévision de niveau 90% et 95% des nombres de tumeurs du groupe expérimental et nombres de tumeurs observées des groupes expérimentaux.

Le graphique ci-dessous représente maintenant pour chaque sexe les mêmes intervalles de prévision, mais avec cette fois-ci le nombre de tumeurs du groupe contrôle :

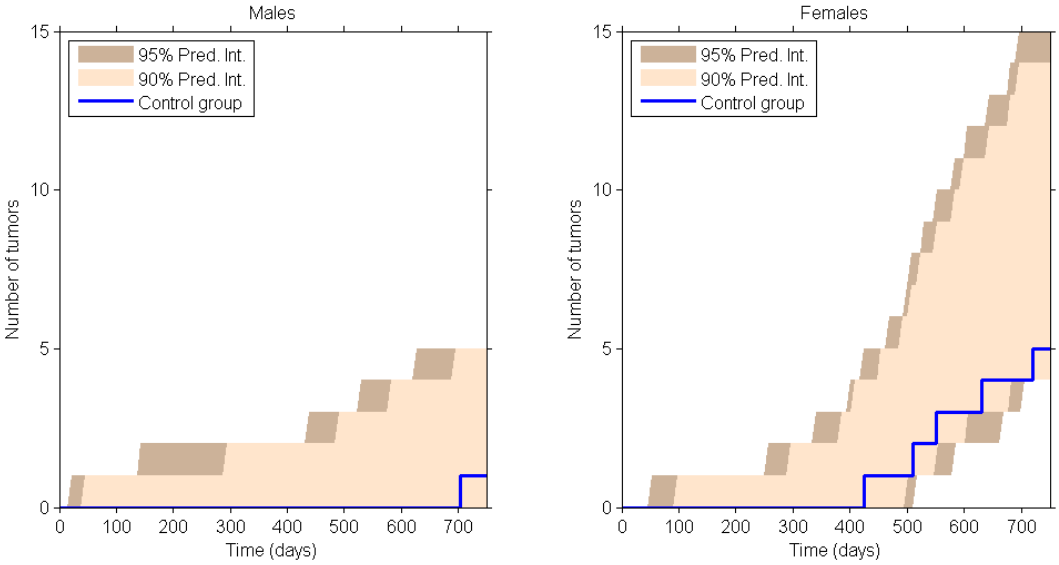

Fig. 13 Intervalles de prévision à 90% et 95% des nombres de tumeurs du groupe expérimental et nombres de tumeurs observées des groupes contrôle.

Les courbes de nombres de tumeurs des groupes contrôles sont à l'intérieur de ces intervalles :

- **On ne peut donc conclure à un effet statistiquement significatif du régime sur l'évolution du nombre de tumeurs.**

6. Paramètres biochimiques

48 paramètres biochimiques ont été mesurés lors de l'étude auprès de chacun des 20 groupes. Pour chaque sexe, pour chaque condition expérimentale, une méthode de type « Orthogonal Partial Least Squares Discriminant Analysis » (OPLS-DA) est mise en œuvre pour discriminer le groupe contrôle et le groupe expérimental.

La méthode OPLS-DA est largement utilisée en chimiométrie ou en génomique pour identifier un sous ensemble de variables qui différencient au mieux différents sous-groupes. Elle est tout particulièrement pertinente lorsque le nombre de variables explicatives est grand devant le nombre d'observations. De plus, la méthode OPLS-DA permet de construire un modèle prédictif, qui, pour un jeu de variables explicatives donné, fournit les probabilités d'appartenir à chacun des sous-groupes considérés.

Le choix de cette méthode et son utilisation par les auteurs dans ce contexte appellent plusieurs commentaires:

1. Ce type de méthode est réputé pour « sur-ajuster » les données observées lorsque le nombre de variables explicatives est grand devant le nombre d'observations (ce qui est le cas ici). En effet, on pourra toujours trouver un modèle défini par 48 paramètres qui séparera parfaitement 2 groupes de 10 sujets, et ce, quels que soient les groupes ! Pour valider le modèle obtenu (*i.e.* s'assurer qu'il possède bonnes propriétés prédictives), on peut utiliser
    - un échantillon test afin de s'assurer que le modèle ajusté sur un échantillon d'apprentissage conserve de bonnes propriétés prédictives sur de nouvelles données qui n'ont précisément pas été utilisées pour la construction du modèle,
    - des méthodes de validation croisées (ce qui revient à faire jouer aux différentes données alternativement le rôle d'échantillon d'apprentissage et d'échantillon test.

    Les auteurs de l'étude ne présentent aucun critère de validation des modèles obtenus à partir des groupes expérimentaux de 10 rats. Ces modèles ne peuvent donc être utilisés en prédiction.

2. Un modèle donné est défini par 48 paramètres : on construit donc ici 18 modèles définis chacun par 48 paramètres. Autant de modèles construits avec aussi peu de données n'ont que très peu d'intérêt : ces modèles sont peu stables et offrent un pouvoir prédictif très limité. Il aurait été plus pertinent de construire un modèle unique intégrant des effets régimes et sexe ainsi que d'éventuelles interactions régime-sexe (voir d'éventuelles relations dose-effet non linéaires). L'avantage d'un modèle unique est qu'il est construit à partir de l'ensemble des données. On évite ainsi une sur-paramétrisation du modèle qui gagne en stabilité et en pouvoir prédictif.

3. L'utilisation de ces méthodes suppose implicitement une distribution symétrique (proche autant que possible de la distribution normale) des variables explicatives. Il est connu que des paramètres tels que les paramètres biochimiques ont une distribution asymétrique et qu'une transformation préalable

est nécessaire pour les rendre le plus « gaussien » possible. Il est par exemple classique d'utiliser certains log-paramètres plutôt que les paramètres d'origine. Un rapport de l'ANSES[4] suggère d'utiliser une transformation Box-Cox pour chaque paramètre, le paramètre de puissance étant le même pour les différents groupes, différents paramètres de position caractérisant chaque groupe.

4. Calculer des intervalles de confiance pour chaque paramètre n'est pas pertinent lorsque de nombreux paramètres sont utilisés. En effet, les éventuelles corrélations entre paramètres et l'aspect multidimensionnel sont totalement ignorés. Il faudrait donc
    - pouvoir calculer des ellipses de confiance, afin de prendre en compte d'éventuelles corrélations entre paramètres,
    - corriger les intervalles de confiance afin de contrôler de façon correcte le risque de première espèce (intervalles multiples).

5. Ces méthodes de classification restent très empiriques et ne sont pas adaptés dans un contexte inférentiel, pour lequel il est nécessaire de calculer des degrés de signification (p-values) et/ou construire des intervalles de confiance. En effet, les lois des statistiques utilisées sont très mal connues et les techniques de type bootstrap ou jack-knife mises en œuvre pour calculer des intervalles de confiance n'ont pas de justification rigoureuse.

Au-delà du choix discutable de la méthode OPLS-DA dans le cadre de cette étude, une erreur méthodologique vient remettre en question les résultats présentés. En effet,

i) 18 comparaisons entre groupes expérimentaux et groupes contrôles sont proposés. Le groupe des femelles nourries avec un régime NK603 33% est celui qui présente le plus de différences : c'est donc ce groupe que les auteurs choisissent de présenter.

ii) 48 paramètres sont comparés. Les paramètres biochimiques présentant le plus de différences (entre le groupe femelle NK603 33% et le groupe contrôle) sont les paramètres *Na, Cl, U.Cl, U.N* tandis que les 2 hormones qui présentent le plus de différences sont *Testosterone* et *Estradiol* : ce sont donc ces 6 paramètres que les auteurs choisissent de présenter.

[4] Recommandations pour la mise en œuvre de l'analyse statistique des données issues des études de toxicité sub-chronique de 90 jours chez le rat dans le cadre des demandes d'autorisation de mise sur le marché d'OGM
http://www.afssa.fr/Documents/BIOT2009sa0285Ra.pdf

Il est alors attendu qu'en sélectionnant à la fois le groupe et les 6 paramètres qui présentent le plus de différences, des différences entre groupe expérimental et groupe contrôle seront visibles. Une telle approche ne permet pas

- **d'expliquer les différences observées par la différence de régime administré.**
- **de rejeter l'hypothèse que c'est la variabilité naturelle des données (dues aux fluctuations d'échantillonnage) ainsi que le critère de sélection des paramètres (les paramètres présentant le plus de différences) qui expliquent les différences observées.**

Conclusion

Le protocole et les outils statistiques utilisés souffrent de graves lacunes et faiblesses méthodologiques qui ne permettent absolument pas de soutenir les conclusions avancées par les auteurs.

i) **Une analyse statistique rigoureuse des résultats obtenus lors de cette étude ne met en évidence**
   - **aucune différence statistiquement significative de la mortalité des rats dans les groupes contrôle et expérimentaux,**
   - **aucune différence statistiquement significative des nombres de tumeurs dans les groupes contrôle et expérimentaux.**

ii) **La méthodologie statistique employée pour l'analyse des paramètres biochimiques est inadéquate et ne permet pas de conclure à l'existence de différences statistiquement significatives entre les groupes traités et témoin.**