QUELQUES COMMENTAIRES D'ORDRE STATISTIQUE SUR

Contents lists available at SciVerse ScienceDirect

Food and Chemical Toxicology

journal homepage: www.elsevier.com/locate/foodchemtox

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

Gilles-Eric Séralini [a,*], Emilie Clair [a], Robin Mesnage [a], Steeve Gress [a], Nicolas Defarge [a], Manuela Malatesta [b], Didier Hennequin [c], Joël Spiroux de Vendômois [a]

[a] University of Caen, Institute of Biology, CRIIGEN and Risk Pole, MRSH-CNRS, EA 2608, Esplanade de la Paix, Caen Cedex 14032, France
[b] University of Verona, Department of Neurological, Neuropsychological, Morphological and Motor Sciences, Verona 37134, Italy
[c] University of Caen, UR ABTE, EA 4651, Bd Maréchal Juin, Caen Cedex 14032, France

1

Marc Lavielle

Inria Saclay
&
Laboratoire de Mathématiques, Université Paris-Sud Orsay

Membre du Conseil Scientifique du Haut Conseil des Biotechnologies

[1] http://dx.doi.org/10.1016/j.fct.2012.08.005

1. INTRODUCTION

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 :

	Mâles	Femelles
Contrôle	10	10
Maïs OGM 11%	10	10
Maïs OGM 22%	10	10
Maïs OGM 33%	10	10
Maïs OGM 11% + Roundup	10	10
Maïs OGM 22% + Roundup	10	10
Maïs OGM 33% + Roundup	10	10
Roundup 50 ng/l	10	10
Roundup 400 mg/l	10	10
Roundup 2,25 g/l	10	10
	100	**100**

L'étude a duré 2 ans durant lesquels plusieurs analyses ont été effectuées :

- Une analyse de mortalité,
- Une étude de pathologies anatomiques,
- Une analyse de paramètres biochimiques.

Le corps de l'article se limite essentiellement à une description des résultats obtenus lors de cette étude. Les remarques concernant cette partie de l'article concernent le choix des différences observées mis en avant. En effet, une telle analyse descriptive ne devrait pas soulever de remarques particulières si les auteurs se contentaient de décrire de façon objective ce qu'ils ont observé chez les différents groupes de 10 rats (courbes de mortalité, pathologies anatomiques,…). Ils ont malheureusement parfois tendance à sélectionner soigneusement quelles comparaisons présenter. 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 expérimentaux de mâles

ayant reçu les plus grandes dose de NK603 et/ou de Roundup présentent des taux de mortalités de seulement 10% ou 20% (1 ou 2 rats morts). Pourquoi ne pas avoir décrit cette différence ?

Et pourquoi ne montrer que des photos de rats issus des groupes expérimentaux ? les tumeurs des rats des groupes contrôles ne sont-elles pas semblables ? Là encore, comme pour les courbes de mortalité, une présentation partielle (et partiale) des résultats ne rend pas compte de l'expérience telle qu'elle a réellement menée.

Le contenu de l'article devient franchement critiquable lorsque les auteurs sortent du domaine purement descriptif des observations en cherchant à expliquer les résultats obtenus et à les généraliser. On peut ainsi lire en conclusion :

- *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.*

De telles affirmations ainsi formulées et ne laissant la place à aucun doute, devraient impérativement 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.

<u>Rappel</u> : nous sommes dans un environnement incertain !

Ce n'est pas parce que seulement 2 rates parmi les 10 du groupe contrôle sont mortes à la fin de l'étude, contre 6 du groupe OGM 22% que l'on peut conclure que le risque de mourir dans les 2 ans pour une rate est 3 fois plus élevé si elle est nourrie avec un régime contenant 22% de NK603.

Le rôle de la statistique inférentielle consiste précisément à évaluer les incertitudes et les probabilités de se tromper en concluant à la présence ou à l'absence d'effets. Il est regrettable que les auteurs aient totalement négligé cet aspect de la statistique, tout en s'autorisant à des surinterprétations non justifiées de leurs résultats expérimentaux.

En suivant la démarche des auteurs (qui consiste à généraliser directement ce qui est observé sur un échantillon réduit à l'ensemble de la population), pourquoi ne pas repris la différence observé entre mâles nourris au NK603 33% et le groupe contrôle pour conclure qu'une forte dose de NK603 réduit le taux de mortalité chez les mâles ? (tout ceci est bien sûr ironique… personne n'oserait remettre en question le fait que cette différence n'est due qu'aux fluctuations d'échantillonnage… tout comme les autres différences observées…)

Le protocole et les outils statistiques utilisés souffrent de graves lacunes et faiblesses méthodologiques qui remettent totalement en question les conclusions avancées par les auteurs.

Une analyse statistique rigoureuse des résultats obtenus lors de cette étude ne met en évidence

- **aucune différence significative de la mortalité des rats dans les groupes contrôle et expérimentaux,**
- **aucune différence significative des paramètres biochimiques.**

2. LE PROTOCOLE EXPERIMENTAL

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.

Il en résulte un manque de puissance puisqu'il devient très difficile de savoir 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.

> 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.

Ce n'est pas du ressort du statisticien (mais au toxicologue) de juger de la pertinence des régimes définis pour cette étude. En admettant que l'étude exige effectivement de constituer 9 groupes expérimentaux pour les mâles et pour les femelles, il aurait fallu

réaliser un calcul préalable du nombre de sujets nécessaires pour mettre tel ou tel effet en évidence. Cela n'a clairement pas été fait.

D'autre part, l'absence totale de plan d'analyse statistique est surprenante, et là encore fortement critiquable. 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.

3. ANALYSE DE SURVIE

L'analyse de survie telle qu'elle est réalisée dans cet article est très incomplète puisqu'elle se limite à une représentation graphique des courbes de mortalité dans chaque groupe (nombre de rats morts en fonction du temps).

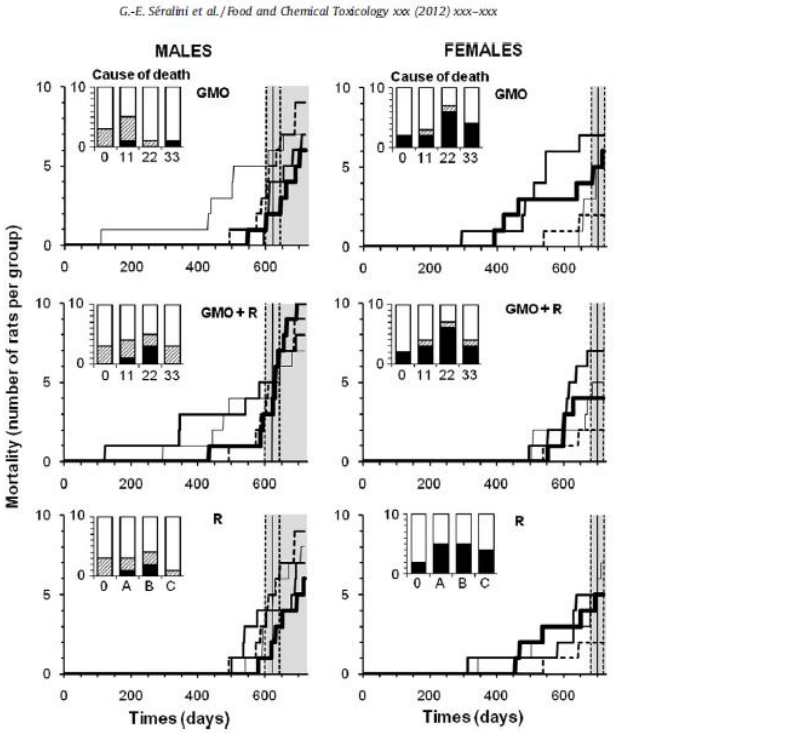
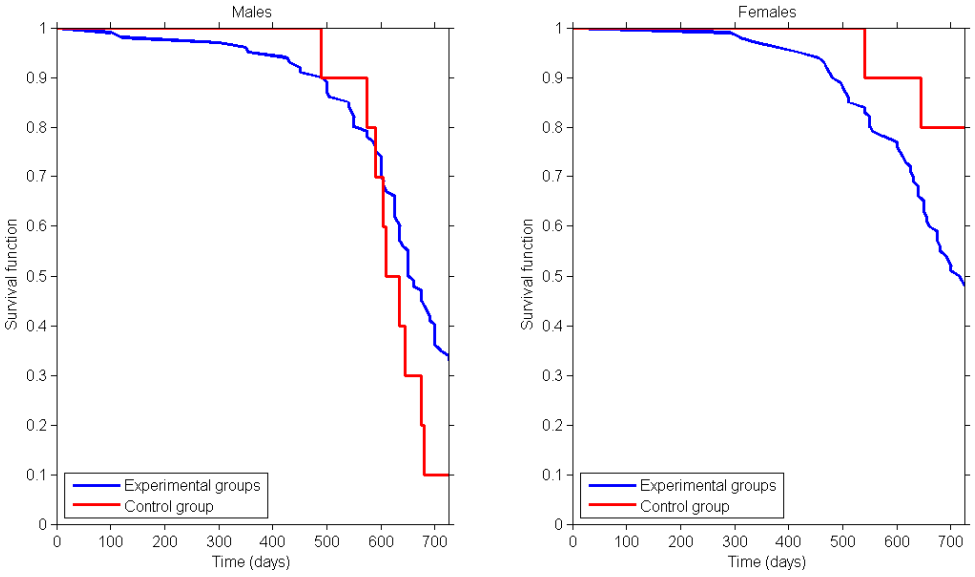


Fig. 1. Mortality of rats fed GMO treated or not with Roundup, and effects of Roundup alone. Rats were fed with NK603 GM maize (with or without application of Roundup) at three different doses (11, 22, 33% in their diet; thin, medium and bold lines, respectively) compared to the substantially equivalent closest isogenic non-GM maize (control, dotted line). Roundup was administrated in drinking water at 3 increasing doses, same symbols (environmental (A), MRL in agricultural GMOs (B) and half of minimal agricultural levels (C), see Section 2). Lifespan during the experiment for the control group is represented by the vertical bar ± SEM (grey area). In bar histograms, the causes of mortality before the grey area are detailed in comparison to the controls (0). In black are represented the necessary euthanasia because of suffering in accordance with ethical rules (tumors over 25% body weight, more than 25% weight loss, hemorrhagic bleeding, etc.); and in hatched areas, spontaneous mortality.

Ces courbes décrivent les mortalités observées. L'examen de ces courbes ne met pas en évidence d'effet flagrant liant le régime et la mortalité chez les groupes expérimentaux. On peut donc, pour chaque sexe, regrouper les groupes expérimentaux et construire une unique courbe de survie (probabilité d'être en vie au cours du temps) que l'on compare à la courbe de survie du groupe contrôle.

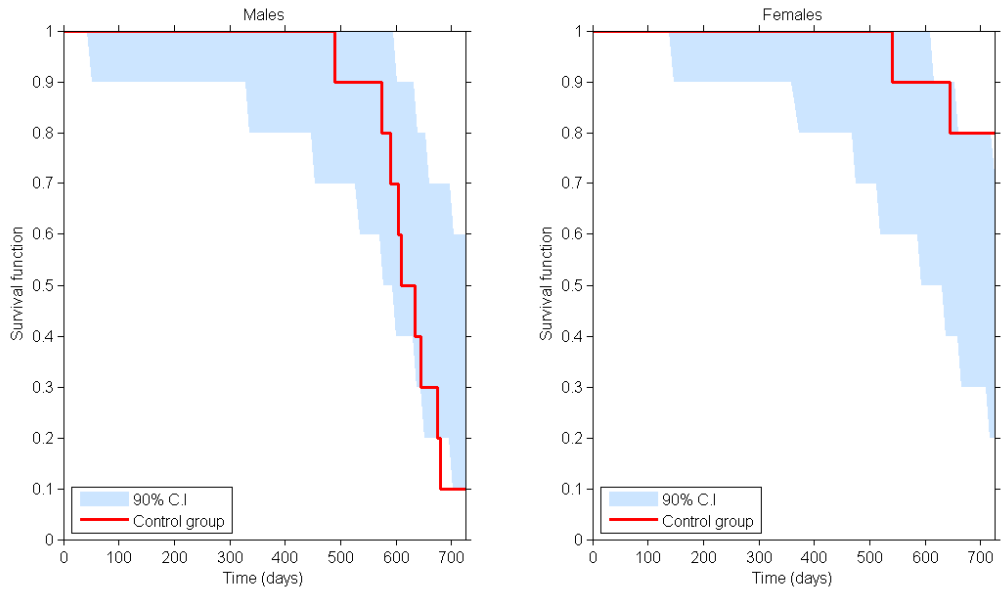


D'un point de vue strictement descriptif, les males du groupes contrôle survivent plus longtemps que ceux du groupe expérimental, mais présentent une mortalité supérieure en fin d'étude. Chez les femelles, par contre, la mortalité dans le groupe contrôle est toujours moins élevée que dans le groupe expérimental.

> On est naturellement tentés d'extrapoler les résultats expérimentaux observés en se demandant, par exemple, si, pour chaque sexe, l'espérance de vie d'un rat nourri avec un régime OGM ou Roundup est réduite par rapport à un rat nourri sans OGM ni Roundup.
>
> Ce n'est pas un simple examen des courbes observées qui permet de répondre à une telle question ! Il faut mettre en œuvre un test statistique rigoureux qui prend en compte la variabilité des individus et donc des courbes de survie.

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

Le graphique ci-dessous représente pour chaque sexe un intervalle de confiance (90%) pour la survie d'un groupe de 10 rats tests[2].



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

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

[2] Un intervalle de confiance peut être facilement construit par simulation. On utilise la courbe de survie du groupe expérimental (bleue) pour simuler un très grand nombre (10 000 par 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 confiance à 90% en calculant à chaque instant les quantiles empiriques d'ordre 5% et 95% des 10 000 courbes de survie.

Une telle conclusion peut sembler surprenante pour les femelles qui, d'après le graphique 1, semblent survivre plus longtemps dans le groupe contrôle que dans les groupes expérimentaux.

Cette conclusion s'explique par la faible puissance[3] des tests statistiques que l'on peut mettre en œuvre avec des groupes contrôle de seulement 10 rats : o**n ne peut rejeter l'hypothèse que cette faible mortalité n'est finalement due qu'aux aléas de l'échantillonnage**.

Ceci illustre à quel point une analyse préalable de puissance et un calcul de nombre de sujets nécessaires sont indispensables pour se donner les moyens de répondre à une question posée.

L'analyse des pathologies anatomiques souffre des mêmes insuffisances que l'analyse de mortalité. On ne peut conclure sur la base des données observées à une différence significative entre groupes contrôles et groupes expérimentaux.

4. PARAMETRES 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 différencie 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

[3] Ce manque de puissance dû à un très faible effectif des groupes contrôle (10 rats) pourrait être compensé par l'introduction d'informations à priori sur le comportement attendu des groupes contrôle. En effet, différentes études de mortalité et de cancérogénèse existent sur la souche choisie pour cette étude (Sprague-Dawley) et pourraient 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 pourrait donc être introduit en utilisant cette information a priori. Au contraire, l'information apportée par les groupes contrôles n'est pas biaisée puisque 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. Un compromis optimal « biais-variance » peut être obtenu en combinant de façon optimale l'information « a priori » fournie par des expérimentations antérieures, et l'information fournie par les données de l'expérience.

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 et l'utilisation de cette méthode dans ce contexte sont surprenants et critiquables pour de multiples raisons :

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 ont omis de valider les modèles obtenus à partir des groupes expérimentaux de 10 rats. Ils 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. On peut se demander l'intérêt de cette approche ! Pourquoi ne pas plutôt avoir essayé de construire un modèle unique intégrant des effets régimes et sexe ainsi que d'éventuelles interactions régime-sexe.

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 « normal » possible. Nous avons suggéré[4] 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] 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

4. 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.

5. 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 α (intervalles multiples).

Au-delà du choix discutable de la méthode OPLS-DA dans le cadre de cette étude, une erreur méthodologique grave 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.*

Il est clair et 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.

- **Néanmoins, on ne peut pas expliquer les différences observées par la différence de régime administré.**

- On ne peut 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.