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

G.E. Séralini *et al.*

Food and Chemical Toxicology, 2012

Statistical Analysis[1]

Scientific Committee of the High Council of Biotechnology

2012, October 22nd

1. The experimental protocol

A sample of 200 rats including 100 males and 100 females was randomized into 20 groups of 10 rats of the same sex. Within each group, rats received the same diet.

For each sex, only one control group of 10 rats was used. It is therefore uniquely this same group of 10 rats that is systematically compared to the 9 experimental groups of the same sex.

This results in a lack of statistical power, meaning that it is difficult to establish whether differences observed when comparing the control with the 9 other groups are due to diet, or merely to natural variation within the control group. Thus, if a certain parameter is particularly high for the control group, all of the differences between the experimental and control groups will tend to have the same trend, showing a decrease in the parameter while not necessarily implying that the effect is due to diet. For example, the authors conclude without advising precaution in §3.3 that "*Creatinine or clairance decreased in urine for all treatment groups in comparison to female controls (Table 3)*", though it is likely that the particularly high values of these parameters in the control group are the reason for these differences.

No initial calculation is provided to suggest the number of subjects required to detect a biologically significant effect. Such a calculation would have been particularly useful for evaluating the amount of information, for example concerning survival and number of tumors, which we might expect to obtain with

[1] English version by Kevin Bleakley, Inria Saclay

the chosen protocol. It would have therefore been possible to quantitatively consider the potentially problematic choice of using a strain of rat that naturally develops tumors with a high probability. In effect, the higher the risk of naturally developing tumors, the more animals per group required to exhibit a significant increase in the number of tumors due to diet. Consider for example two strains of rat, A and B, for which the risk of naturally developing a tumor within a given period of time is 10% for A and 60% for B. The tables below show the number of rats required to suggest an increase in the number of tumors of 10%, 20% or 30% when the type I error α and type II error β are both equal to either 5% or 10%. We see that using strain B requires more rats that strain A, independent of the given error risks, when the increase to be revealed is less than 30%.

		Rat strain	
		A p=10%	**B** p=60%
Effect to be shown	+ 10%	135	248
	+ 20%	41	60
	+ 30%	24	24

α=β=5%

		Rat strain	
		A p=10%	**B** p=60%
Effect to be shown	+ 10%	86	156
	+ 20%	25	36
	+ 30%	15	15

α=β=10%

Table 1. Number of rats required in an experimental group in order to suggest a given increase in the number of tumors for two strains of rat A and B (risk of naturally developing tumors in a given time period is respectively 10% and 60%) for two given type I and II error risks[2].

More generally, the article does not mention the statistical protocol. The authors appear to have undertaken their statistical analysis as a function of the results obtained, entirely contradictory to elementary rules of good statistical practice. In effect, the statistical significance of an observed difference in a given parameter is not the same when the parameter has been selected *a priori* (before obtaining results) or *a posteriori* (from among the parameters exhibiting the largest differences). The authors give in Figure 5-B the 4 biochemical parameters and 2 hormones, which exhibit the largest differences, within the group that shows the most differences. This choice has been made *a posteriori*, i.e., after the results were obtained. It is statistically expected that some of the 18x48=864 comparisons will provide differences that appear significant. Presenting these partial results in this way may mislead a non-specialist in multiple comparison (here 864) statistics to conclude that the observed differences are due to the difference between experimental and control conditions.

[2] The type I error α is the probability of concluding that there is an increase even though there is not. The type II error is the probability of not detecting an increase even though there is one.

2. Descriptive analysis of results

The body of the article is essentially limited to a description of the results obtained for the different groups of 10 rats (survival curves, anatomic pathology, etc.).

The author's comments partially present what was observed. For example we read in §3.1 that "*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).*" While certain experimental groups of males had in effect a mortality rate of 50% (5 deceased rats) after 600 days, the experimental groups of males that received the largest dose of NK603 and/or Roundup had mortality rates of only 10% (1 deceased rat). This observed difference is not mentioned. Also, the choice of looking at the mortality rate at 600 days is totally arbitrary: the mortality rate of 30% in the male control group at 600 days becomes 50% at around 620 days.

Furthermore, it is difficult to accord a statistical meaning to the various photos (rats, organs, tumors, etc.) because only some rats are shown. We therefore must ask whether the selected rats are representative of their group. Indeed, photos for the control groups should also be provided.

3. Statistical inference

Inferential statistics allow us to evaluate the incertitude and probability of making a mistake when making conclusions about the presence or absence of effects; that is, if the observed differences can be explained by a change in diet or are simply due to random fluctuations in the sampling. Said another way, the natural question to ask here concerns reproducibility of the results: if we repeat the same experiment under the same conditions, what are the chances of obtaining similar results?

As for the survival analysis and the counting of the number of tumors, the authors have entirely ignored this statistical aspect, while proposing unsubstantiated interpretations of their experimental results. On pages 8-9, we read:

- ***All treatments in both sexes enhanced large tumor incidence*** *by 2–3-fold in comparison to our controls…*
- *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.***

Not a single statistical argument can be found in this article to suggest a cause-effect relationship of this type. There is not a hint of statistical analysis that would suggest a statistically significant difference in survival and number of tumors between the experimental and control groups.

As for the biochemical parameters, we may read in the article's 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.*

Such affirmations merit rigorous justification and validation. Here, it is absolutely impossible to conclude with certainty the toxicity of NK603 on such a restricted set of data.

4. Survival analysis

4.1 Survival

The authors explain in §3.1 that *"Control male animals survived on average 624 ± 21 days, whilst females lived for 701 ± 20"*. These values are the result of an incorrect calculation because the data is censored (we do not know when the animals still alive at end of study would have died naturally since they were euthanized). Instead, what has been calculated is the empirical mean and standard deviation of the uncensored observed values joined with the censored values for still-alive rats (as if a rat still alive at T=720 days is considered dead at T=720 days). The results given are therefore inexact because this procedure introduces a bias by underestimating the average date of death and clearly also the standard error of the estimator. To have chosen not to consider everything that occurs after 624-21=603 days is therefore not justified because this value comes from an incorrect calculation.

Correct calculation of the survival distribution for different groups requires the introduction of a parametric model, but the use of such an approach is constrained, given the limited amount of data per group. For example, if we fit a Gaussian model for the survival time of the males, the estimated mean and standard deviation are respectively 626 and 68 days. For the females, it is 892 and 206 days! It is not correct to proceed as the authors have done and calculate the standard error for the mean merely by dividing the standard deviation by √10, as would be done for uncensored Gaussian variables. Due to censoring, the distribution of the estimator of the mean is much more spread out and asymmetric, meaning that the use of the standard error for calculating confidence intervals is not meaningful.

4.2 Comparisons between experimental and control groups

Survival analysis in this study was limited to graphical representation of mortality curves in each group (number of deceased rats as a function of time), shown in the following figure.

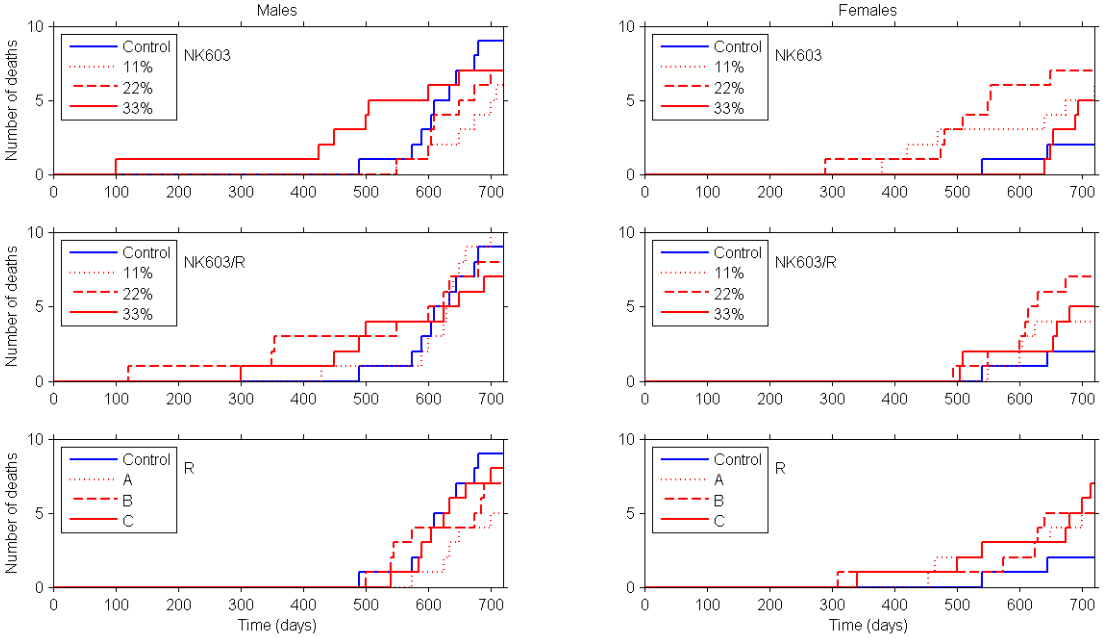

Fig. 1 Mortality curves for the 18 experimental and 2 control groups.

A visual inspection of these observed mortality curves does not provide the slightest evidence of differences between experimental and control groups.

Numerous statistical techniques exist for comparing survival/mortality curves. To begin with, we might consider the 18 possible comparisons between the experimental and control groups. The Wilcoxon rank-sum test is a nonparametric test that allows comparison of the rank statistics of two samples.

In this way we can test whether the rats from a given experimental group have a tendency to die earlier than the control group. For example:

H_0 "the NK603 11% diet has no effect on the survival of female rats"

vs

H_1 "the NK603 11% diet leads to decreased survival time for female rats."

The test statistic is defined as the sum of the ranks of the control group. For each of the 18 comparisons, this test statistic can be calculated and compared with a prediction interval obtained under the null hypothesis. The statistical significance can then be calculated for each of the 18 tests:

.

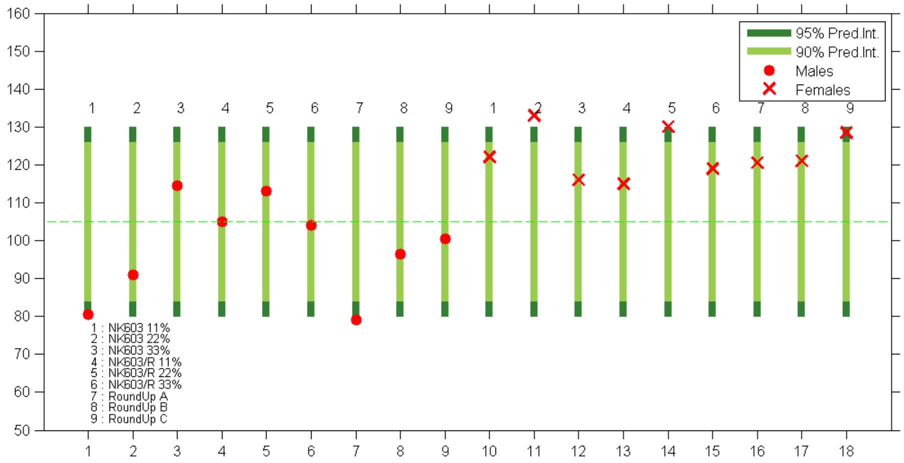

Fig. 2a 90% and 95% prediction intervals of 18 test statistics along with the observed values of these test statistics.

However, these prediction intervals do not take into account the multiple comparison testing involved. Rather than putting into practice an overly conservative test (which has a tendency to systematically not reject the null hypothesis), we can estimate by either simulation or permutation the probability distribution of the 18 statistics used for this test. The following figure shows the prediction intervals of the 18 test statistics put in decreasing order (intervals estimated by simulation here). The 18 test statistics are all within the corresponding 90% prediction intervals:

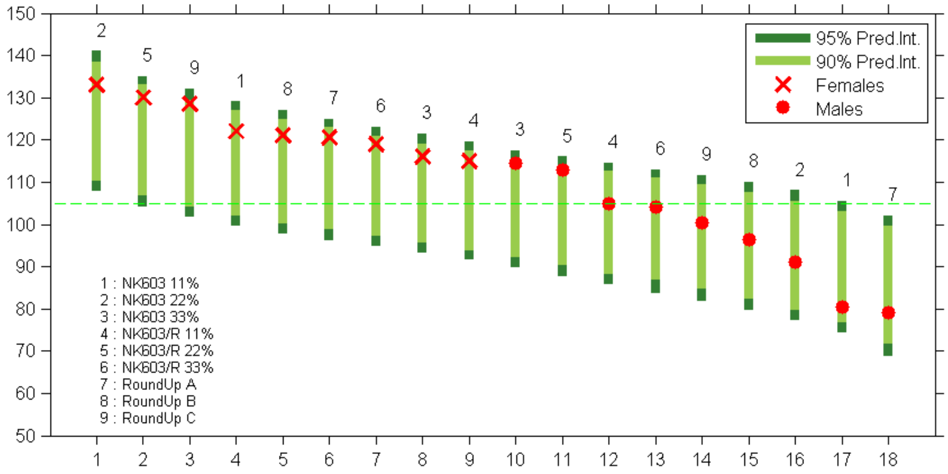

Fig. 2b Intervalles de prévision de niveau 90% et 95% des 18 statistiques de tests ordonnées par ordre décroissant et valeurs observées de ces statistiques de tests.

We can also estimate the statistical significance of each comparison like an empirical quantile. The table shows for each of the 18 tests the test statistic, *i.e.,* the sum of ranks of the control group (under the null hypothesis, the expected value of this statistic is (1+2+…+19+20)/2=105), the statistical significance of the corresponding rank-sum test, and the adjusted statistical significance that takes into account the fact that multiple tests were performed (by simulation or permutation). Groups that have a tendency to die before the control group (test statistic of over 105) are in red; the other ones are in blue:

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 The 18 test statistics (Wilcoxon rank-sum test) and the associated statistical significances.

The group that exhibits the largest difference with respect to survival is the female group with the diet of 22% untreated NK603 corn. The statistical significance is 1.1% (i.e., the probability of obtaining a test statistic greater than or equal to 133 under the null hypothesis is 1.1%). Taking into account the multiple tests undertaken, this probability is 15.8% when estimated by simulation and 16.6% by permutation. Thus, the probability that the largest test statistic among 18 is greater than or equal to 133 under the null hypothesis is around 16%. The table therefore allows us to conclude that:

- **no observed difference between the survival curves of the experimental and control groups is statistically significant.**

Lastly, we can show by simulation that, for example, if the death of 5 rats in an experimental group of 10 occurs before the death of a rat in the control group, the statistical significance of the test is 8%. This drops to 2% (resp. 0.8%) if there are 6 (resp. 7) experimental rats that die before a control rat does.

4.3 The use of reference data

The lack of power, due to the small number of control rats, clearly stops us from being able to formally make conclusions as to the presence or absence of an effect of diet on survival, in particular for the female rats. This lack of power can be compensated by introducing *a priori* information on the expected scenario for the control groups. Indeed, survival data for the SD rat strain are available from the Harlan Company and can be used to add pertinent information to the experimental set-up. Obviously, this data has not been obtained under exactly the same conditions as the present study and bias may therefore be introduced by including this *a priori* information. On the other hand, information coming from the control groups is not biased if all groups were treated under the same conditions, but as we have seen, this information is affected by a large level of variability. A combination of the *a priori* information and the experimental data leads to a good bias-variance compromise.

Here, the data provided by the Harlan Company indicates a 2-year survival rate of 32% for males and 48% for females. For each sex, the number of rats alive after 2 years is a binomial random variable for which we can calculate 90% and 95% prediction intervals[3].

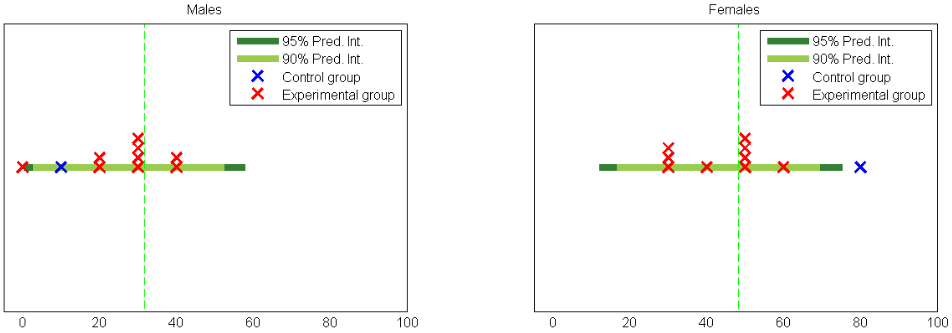

Fig.3 90% and 95% prediction intervals for 2-year survival rate obtained from data from the Harlan Company, and the survival rate observed in the experimental and control groups.

The experimental groups are for the most part distributed within these prediction intervals. With 18 groups, it is entirely normal that one observation be found on the edge of the 95% prediction interval. The 2-year survival rates are entirely in accord with the reference data provided by the company when the rats are raised in normal conditions.

- **The use of reference data provided by the Harlan Company confirms that we cannot explain the differences observed in the survival curves of different experimental and control groups by invoking the effect of diet.**

It can also be noted that the survival rates observed in the control groups are nevertheless relatively far from what the reference values would suggest (the observed proportion of female rats still alive after 2 years is outside the 95% prediction interval). This further emphasizes the fragility of the resulting statistics obtained from such a small number of cases; it is impossible to provide definitive conclusions here.

[3] For binomial variables, only the quantiles of order (*i/n, i*=0, 2,…n) can be directly calculated from the probability distribution. We can obtain any other quantile using linear interpolation.

4.4 Less groups, more power

These results show that it is impossible to consider with sufficient power the 18 possible comparisons between experimental and control groups. The experimental protocol is not at all sufficient for such an ambitious goal.

By limiting the comparisons by regrouping certain groups allows the construction of more powerful and robust tests. For example, we might limit ourselves to test whether mortality in the control group is lower than in the experimental groups in general. We therefore regroup for each sex the experimental groups and construct a single survival curve (probability to be alive with respect to time). The two experimental groups (male and female) are now made up of 90 animals: we can therefore reasonably approximate the true survival functions by the empirical ones obtained from the two samples of 90 rats. These can then be compared to the survival functions of the two control groups.

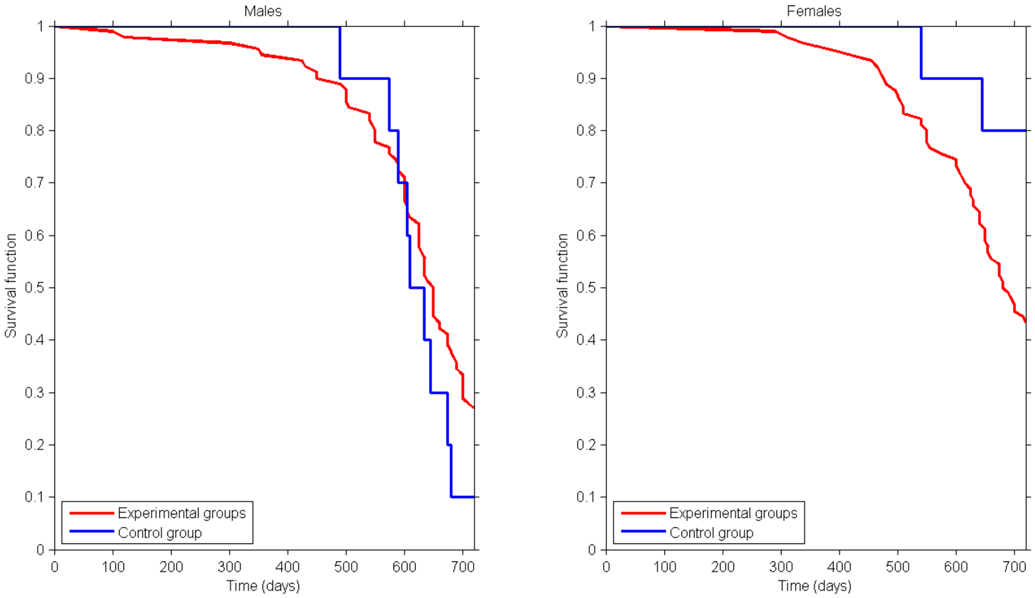

Fig. 4 Survival curves for the experimental and control groups. For each sex, the experimental group is made up of the 9 initial experimental groups.

The question is: could the blue survival curve (observed control group) have been obtained from 10 rats whose survival probability is characterized by the red curve?

A prediction interval for the survival curves of the two experimental groups (male and female) can be constructed easily via simulation. For each sex, we use the survival curve of the experimental group (red) to simulate a very large number (10,000 here) of groups of 10 rats and their dates of death. We can thus construct 10,000 empirical survival curves from the 10,000 groups. From this, we can construct a 1-α% prediction interval by calculating at each instant of time the empirical α/2 and 1-α/2 quantiles of the 10,000 survival curves. We thus obtain a 90% (resp. 95%) prediction interval by calculating the empirical 5% (resp. 2.5%) and 95% (resp. 97.5%) empirical quantiles.

The following figure presents for each sex the survival curves of the 9 experimental groups and the 90% and 95% prediction intervals.

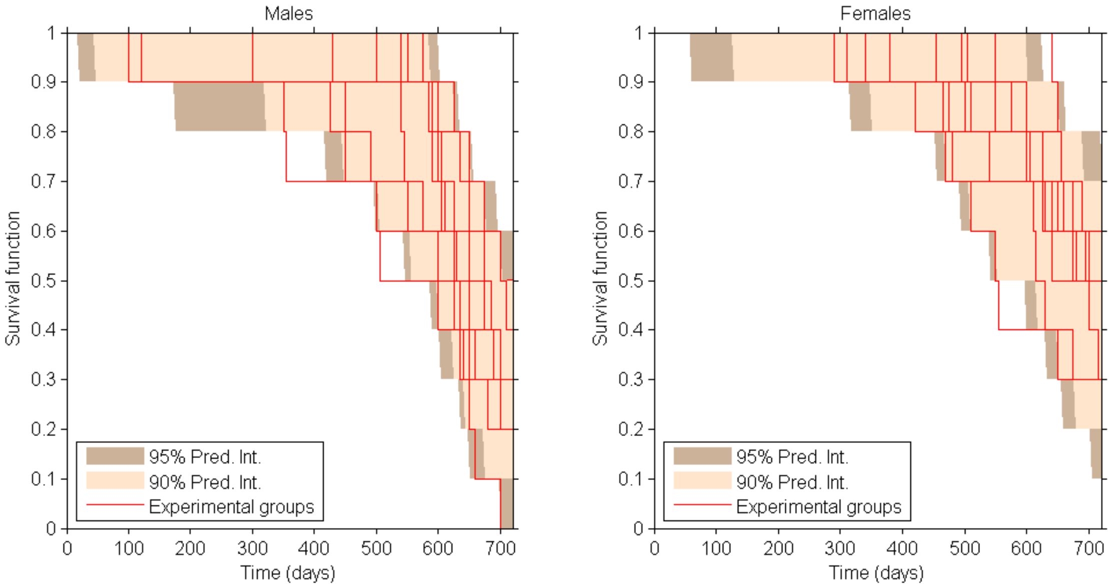

Fig. 90% and 95% prediction intervals for the survival of the combined experimental groups, and the survival curves of the 9 experimental groups seen individually.

The next figure shows for each sex the same prediction intervals, but now with the survival curve of the control groups:

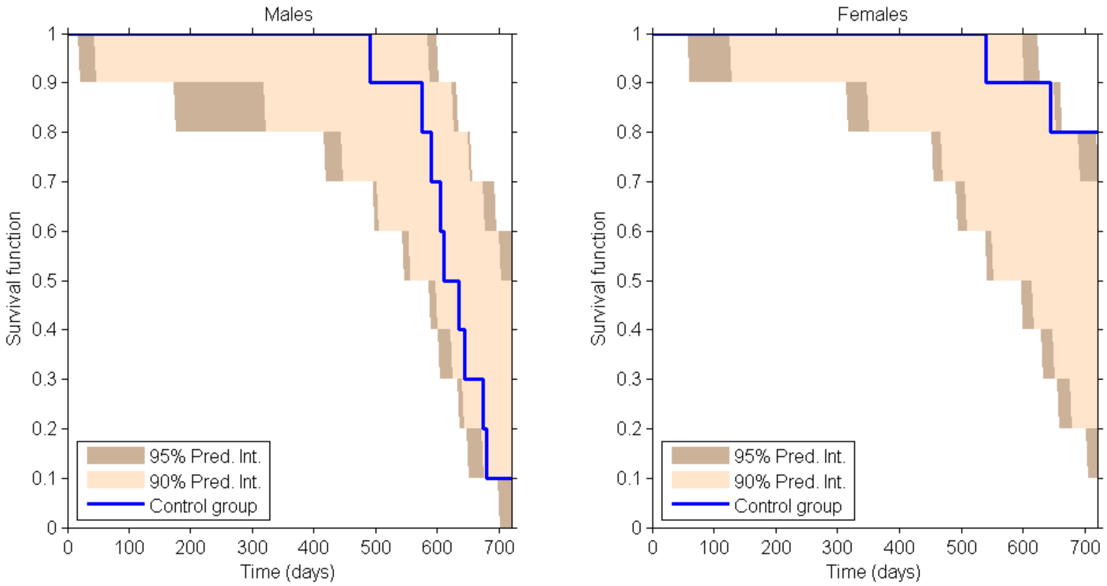

Fig. 6 90% and 95% prediction intervals for the survival of the combined experimental groups, and the observed survival curves of the control groups.

The survival curves of the control groups are essentially inside the prediction bounds.

- **We cannot conclude that there is a statistically significant difference between the survival of the control and experimental rats.**

Again, the reference data provided by the Harlan Company underpin this conclusion because the prediction intervals of the mortality rate at 2 years (for a sample of size *n*=90) contain the observed rates in the experimental groups:

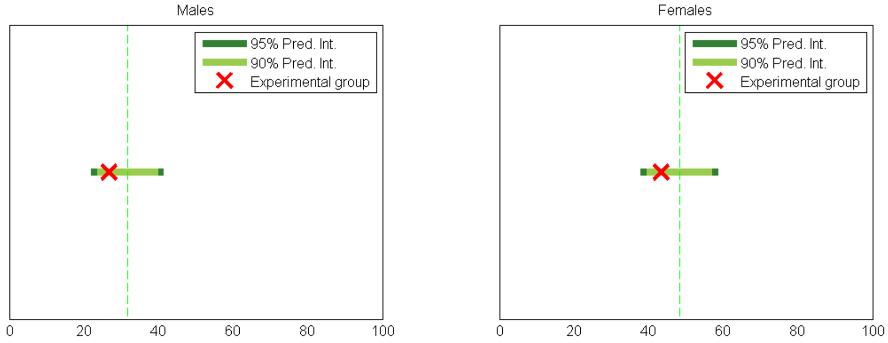

Fig. 7 90% and 95% prediction intervals for the 2-year mortality rate obtained from the Harlan data, and 2-year mortality rates seen in the experimental groups (each of size 90).

The choice to regroup the 9 experimental groups and test whether their distribution is different from the control group is obviously debatable. We might want to implement other tests to test for example whether consumption of a GM corn has an impact on survival, independent of the dose of GM corn and the associated treatment (RoundUp or not). The combined experimental group is then made up of the first 6 experimental groups NK603 and NK603/R, 11%, 22% and 33%), while the control group is just the initial control group (no GM, no GM/R, no R) joined with the 3 groups of rats exposed to RoundUp. The figure below shows that no significant differences exist between these groups for either the males or females. Therefore, we are unable to conclude that there is a significant effect of NK603 corn on rat mortality.

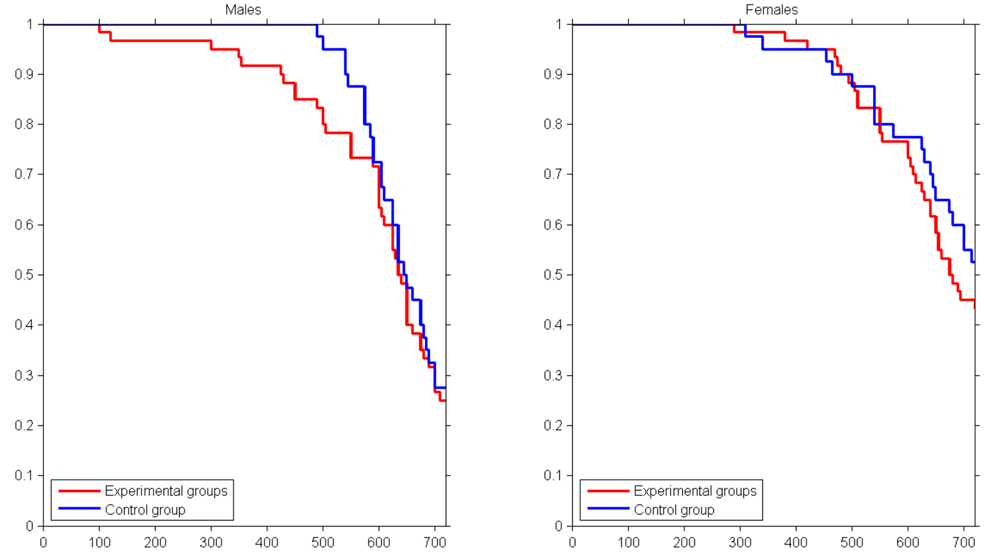

Fig. 8 Survival curves for the experimental and control groups. For each sex, the experimental group contains the 6 initial experimental groups that consumed NK603 corn (with or without RoundUp). The control group contains the initial control group along with the 3 experimental groups that had not consumed NK603 corn.

We can also test whether the absorption of any quantity of RoundUp (in liquid form) has an impact on survival. The experimental group is thus formed of the 3 groups that absorbed RoundUp but not NK603, and the control group is formed by combining all remaining experimental groups with the initial control group. The figure below shows that there is no significant difference in survival between control and experimental groups, both for males and females. We are thus unable to conclude that RoundUp has a significant effect on rat mortality.

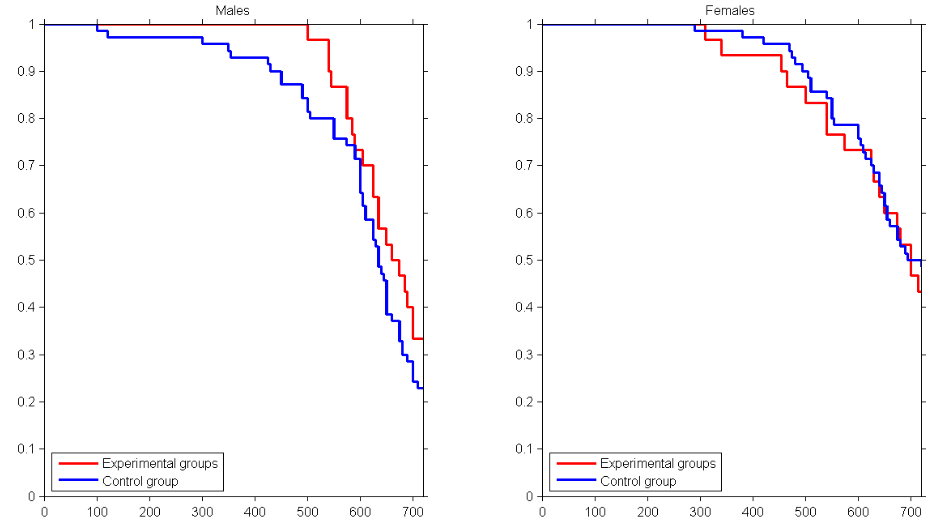

Fig. 9 Survival curves for experimental and control groups. For each sex, the experimental group is made up of the 3 groups that had absorbed RoundUp (but not NK603). The control group is made up of the 6 remaining experimental groups and the initial control group.

In summary, we cannot conclude that there is a statistically significant effect of any treatment (NK603, NK603 treated with R, R) on the survival of rats.

5. Number of tumors

Analysis in the study of the number of tumors was limited to a graphical representation of the number of palpable tumors observed in each group as a function of time:

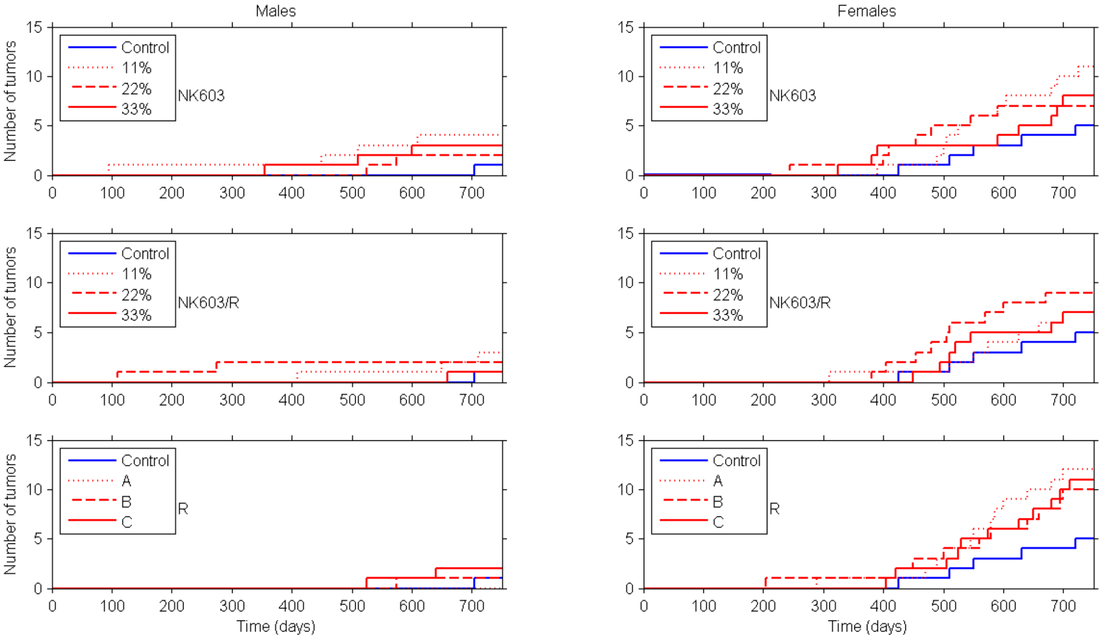

Fig. 10 Evolution of the number of palpable tumors in the 20 groups.

As was the case for the survival curves, simply looking at the curves here does not allow us to make any confident conclusion as to differences between populations. For that, a rigorous statistical test, which takes into account statistical variability and thus variability in the tumor count curves, would need to be implemented.

As before, the experimental protocol is not at all ideal for making the 18 proposed comparisons, and as before, we can for each sex regroup the experimental groups and construct a single curve for the number of tumors, which can then be compared with the control group's one:

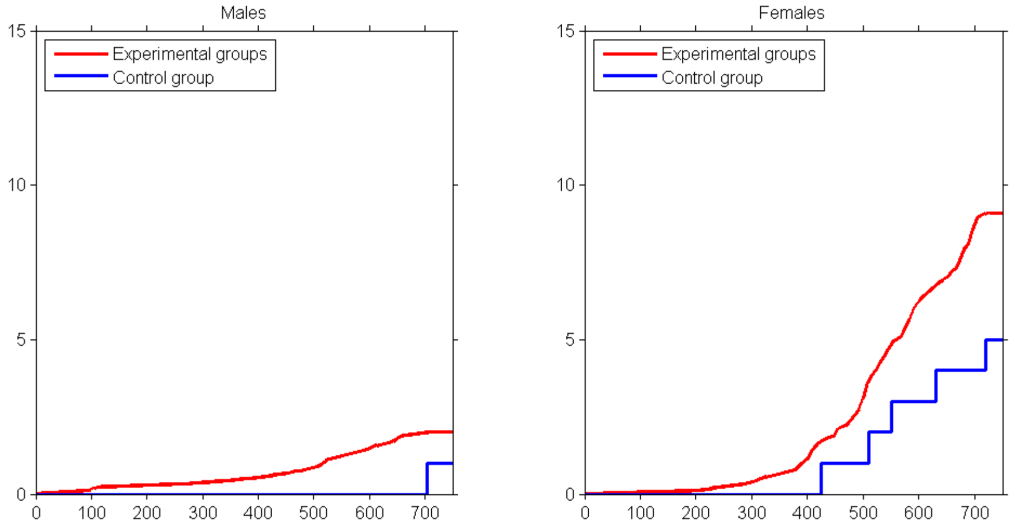

Fig. 11 Number of palpable tumors in the experimental and control groups. For each sex, the experimental group is made up of the combined 9 initial experimental groups.

The natural question is then to ask whether the blue curve for the number of tumors (observed control group) could have been obtained from 10 rats that had been used to create the red curve.

Prediction intervals for the number of tumors in the 2 experimental groups (male and female) can be easily constructed by supposing that for each sex the number of tumors is a nonhomogeneous Poisson process whose intensity is given at each instant of time by the red curve. In this way, 90% (resp. 95%) prediction intervals can be obtained by calculating at each instant of time the 5% (resp. 2.5%) and 95% (resp. 97.5%) quantiles of a Poisson variable.

The following figure shows for each sex, the number of tumors of the 9 experimental groups and the 90% and 95% prediction intervals:

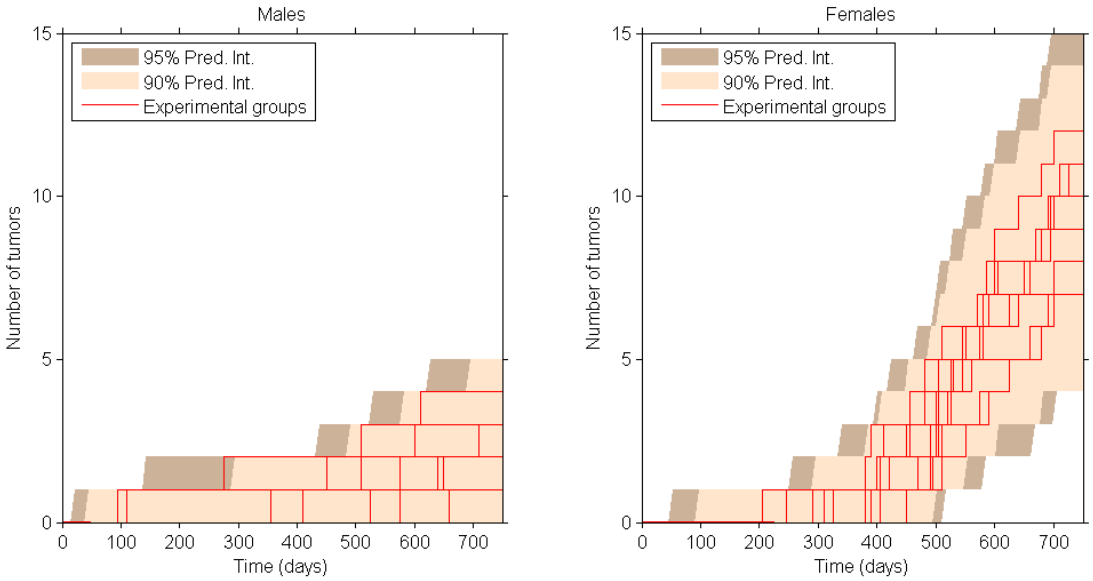

Fig. 12 90% and 95% prediction intervals for the number of tumors in the experimental group, and the evolution of the number of tumors observed in the 9 initial experimental groups.

The following figure then shows for each sex the same prediction intervals, this time along with the evolution of the number of tumors in the control group:

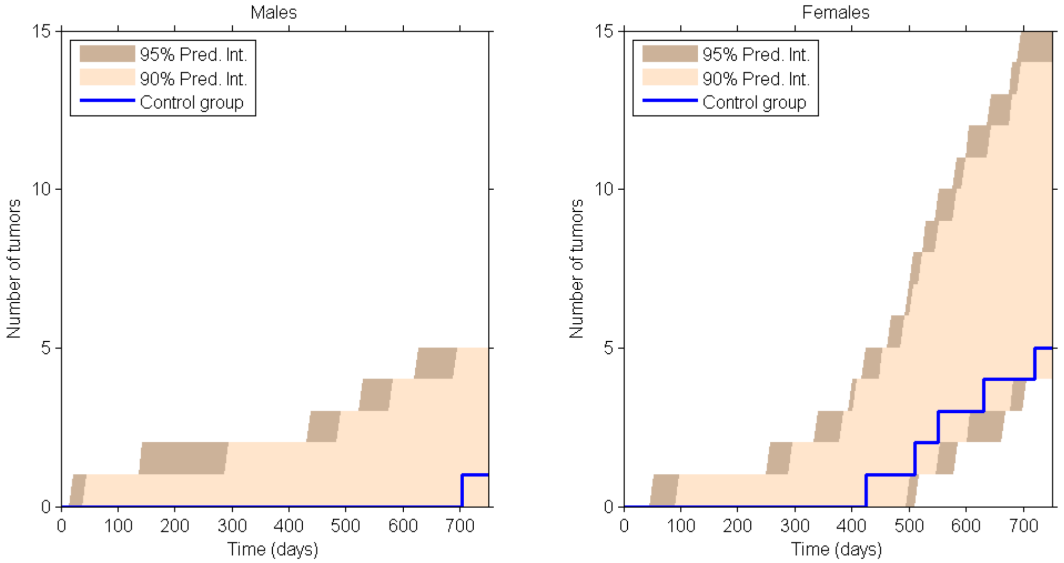

Fig. 13 90% and 95% prediction intervals for the number of tumors in the experimental groups, and the evolution of the number of observed tumors in the control groups.

The curves for the number of tumors in the control groups are inside both prediction intervals:

- **We cannot conclude that there is a statistically significant effect of diet on the number of tumors.**

6. Biochemical parameters

In the study, 48 biochemical parameters were measured for each of the 20 groups. For each sex and for each experimental condition, an "Orthogonal Partial Least Squares Discriminant Analysis" (OPLS-DA) method was implemented to discriminate between the control and experimental group.

The OPLS-DA method is frequently used in chimiometrics and genomics for identifying a subset of variables which can best separate different groups. It is particularly pertinent when the number of explicative variables is large with respect to the number of observations. Furthermore, the method allows construction of a predictive model which, for a given set of explicative variables, can output the probability of belonging to each of the groups under consideration.

The choice of this method and its use by the authors in the present context deserves several comments:

1. This type of method is well-known for "over adjusting" the observed data when the number of explicative variables is large with respect to the number of observations (which is the case here). In effect, it is always possible to find a model defined by 48 parameters that perfectly separates 2 groups of 10 subjects, no matter the groups! To validate the obtained model (i.e., to assure oneself that it possesses good predictive properties), one can use
   - an independent test set, which helps to ensure that the model fitted to the training set retains good predictive properties on new data which has not been previously used to fit the model.
   - cross-validation methods (essentially, different subsets of the data are successively used to play the role of training and test set).

   The study's authors provide no model validation criteria constructed using the experimental groups of 10 rats. These models cannot therefore be used in a prediction framework.

2. A given model is defined by 48 parameters, and 18 models are thus defined. This many models constructed from so little data is far from ideal: such models will be unstable and have extremely limited predictive power. It would have been more useful to construct a single model that integrated the effects of diet, sex, and perhaps their interaction (or even potential nonlinear dose-effect relationships). The advantage of working with a single model is that it would be built using all of the data, thus helping to reduce the over-parameterization of the model, leading to better stability and predictive power.

3. The use of the method implicitly supposes a symmetrical distribution (as close as possible to a Gaussian distribution) of the explicative variables. It is known that parameters such as biochemical ones have asymmetric distributions and that a pre-transformation is necessary to render them as "Gaussian" as possible. For example, it is standard practice to use certain log-parameters rather than the original parameters. A report by ANSES[4] suggests using the Box-Cox transformation for each parameter, the

[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

power parameter being the same for each group meaning that different position parameters characterize each group.

4. Calculating confidence intervals for each parameter is not pertinent when many parameters have been used. In effect, potential correlations between parameters and the multidimensional point-of-view are totally ignored. What is needed therefore is to:
   - be able to calculate confidence ellipses in order to take into account possible correlation between parameters.
   - correct the confidence intervals in order to correctly control the type I error (multiple intervals).

5. These types of classification methods remain quite empirical and are not particularly suited to a prediction context, for which it is necessary to calculate statistical significance (p-values) and/or construct confidence intervals. In effect, the laws of statistics used are poorly understood and methods such as bootstrap and jack-knife, used for calculating confidence intervals, cannot be rigorously justified.

As well as the debatable choice of using the OPLS-DA method in this study, a methodological error calls into question the results presented. In effect,

i) 18 comparisons are proposed between experimental and control groups. The group of females on the diet NK603 33% is the one that shows the largest differences: this is the group that the authors choose to present.

ii) 48 parameters are compared. The biochemical parameters exhibiting the biggest differences (between the female NK603 33% group and the control group) are *Na, Cl, U.Cl, U.N* and the 2 hormones that exhibit the biggest differences are *Testosterone* and *Estradiol*. These are the 6 parameters the authors choose to present.

It is therefore expected that upon selecting both the group and the 6 parameters that exhibit the largest differences, differences between the experimental and control groups will be apparent. Such an approach does not allow us to:

- **propose that the observed differences are caused by diet.**
- **reject the hypothesis that it is natural variability on the data (due to sampling fluctuation) and the parameter selection criteria (choosing those with the largest differences) that explains the observed differences.**

Conclusion

The experimental protocol and the statistical methods used in the article suffer from serious gaps and methodological weaknesses and do under no instance support the conclusions proposed by the authors.

i) A rigorous statistical analysis of the results obtained in this study does not show

- **any statistically significant difference in mortality in rats between the control and experimental groups.**
- **any statistically significant difference in the number of tumors between the control and experimental groups.**

ii) The statistical methodology used to analyse the biochemical parameters is inadequate and does not lead to the conclusion that there are statistically significant differences between the control and experimental groups.