Taking risks – Quantitative severity estimation for behavioral data

Steven R. Talbot
Laboratory Animal Science, Hannover Medical
School (Hannover, Germany)
talbot.steven@mh-hannover.de

ABSTRACT

This is the abstract.

It consists of two paragraphs.

1. INTRODUCTION

Animal studies bear the risk of death. Animals have to be euthanized because severity levels become too grave or they simply die due to direct treatment. In contrast, guidelines like the European Directive 2010/63/EU require that in every animal study stress and exposure to the animals have to be minimized – thereby avoiding potential severity. Researchers are often in a dilemma in which to avoid exceeding harm to animals while studying treatment effects. Monitoring risk in combination with actual severity therefore seems appropriate. However, the terms “risk” and “severity” itself are fuzzy and obtaining objective information from behavioral studies is difficult. There are attempts of better understanding severity as a function of variables but an attempt for objective risk assessment has never been followed to our knowledge.[1]

Measures like scoring systems are considered to be the epitome of severity assessment. They seemingly provide contingent results but they also require highly trained experts. And being human-based they always remain biased as they rely substantially on human assessment. This raises the question for more objective methods that are relatively free of subjective bias. A promising approach is risk monitoring.

Risk can loosely be described as the probability of occurrence (likelihood) times the consequences (severity). The linear relationship of risk calculation is simply

$Risk = Probability * Consequences$

In animal studies this approach is almost never followed numerically, because likelihood estimation is only possible retrospectively, unless experiments are not repeated with the same design or more creative solutions are followed. Instead, a prospective severity is determined based on a qualitative risk assessment matrix (RAM, Figure 2). With this tool it is a rather straightforward process to determine the potential risks and outcomes of undertakings such as animal studies.

Depending on the actual experimental design the terms probability and consequences can mean different things, e.g. vulnerability or loss. But even with this qualitative tool the field of severity assessment remains one of ongoing investigation. Since the diversity of parameters in animal sciences is large, the actual selection of parameters is often limited to the research question. Especially, if variables are not understood well, the actual calculation of risk remains a guessing game.

We have collected a total of four representative data sets from the FOR2591 consortium “Severity Assessment” and will show how body weight data can be used in numerical risk and severity calculation.

Although we exemplarily focus on body weight the actual method can be used for a wide array of behavioral parameters. This will add a new level of flexibility to the system of the rigorous -20% threshold in body weight change as humane endpoint and will also allow untrained scientists to evaluate animals at risk. We will further show how body weight can be used to assess severity within certain constraints, allowing a more objective approach for the identification of animals at risk.

2. MATERIALS AND METHODS

2.1 Software

Mathematical modeling was performed with R feather spray version (R version 3.5.1 (2018-07-02)) on a 64-bit machine with 24 GB RAM. The following packages were used in the calculations: readxl, reshape2, melt, density, kmeans, ecdf.

2.2 Studies

The following studies were used (Table 1). They all represent specific cases that illustrate possible outcomes for risk and severity assessment. Three studies have been published before and are used as case-studies here. Further experimental details can be obtained from the original publications.

[1] Tesfottnote

2.3 Body weight change kernel density calculation

Longitudinal body weight change (%) data with different lengths from each of the four studies were appended resulting in a vector of length 2652. The combined time-free data were then smoothed with a Gaussian kernel resulting in an object with 512 BWC (%) data points. Missing values were omitted. Kernel density data were then used in subsequent calculations.

2.4 Clustering

Output data from the R function "density" was used in the "kmeans" function to calculate five clusters. Seeding was held constant for this step. Clusters where then plotted, color-coded and lines were attached to the outside borders of each cluster. No further optimization was followed.

2.5 Fibonacci retracements

Output data from the R function "density" was halved to 256 data points focusing just on the negative side of the density distribution – assuming equality on both sides of the kernel function. The data points where then divided into the Fibonacci retracement levels: 78.6, 61.8, 50.0, 38.2 and 23.6 % which are derived from the golden ratio. These levels where then color-coded and added to the plot.

2.6 Calculation of sampling distributions

Data were divided into surviving and dead animals independently from actual treatment. Missing entries were omitted. The last BWC (%) value reached before euthanization or death was taken for each dead animal and bootstrapped with 1000 repetitions using the R function "replicate". Results were plotted as histograms which directly reflect the sample distribution of animals at highest possible risk in the analyzed models. The same was repeated for surviving animals but here the sample mean was used for bootstrapping. In case of study D where no dead animals occurred, risk was defined as the minimum reached BWC (%) value. In this case, the lowest reached BWC (%) value in the treatment group was used as well as respective controls.

2.7 Likelihood calculation

Individual study data were reduced to raw body weigh change (%) data using the R "melt" function. A threshold of 5% in either direction of zero was given to allow some breathing space for naturally occurring variance. Using this, data were divided into a low and high fraction. Using the empirical cumulative distribution function "ecdf", the low and high fractions were analyzed separately resulting in two curves ecdf.low and 1-ecdf.high. Both curves show the likelihood of occurring values in the respective study data set. Since there are no relevant negative data for study C, no likelihoods were calculated here. Query points were found manually by calculating the closest data point in the respective likelihood.

2.8 Risk and severity assessment

Studies with likelihood information were used for risk assessment and severity calculation using the general risk formula (1). The probability term was given by the value likelihoods obtained from raw study BWC (%) value distributions. Risk was derived from the bootstrapped sample distributions of the dead animal data or, if no dead animals occurred as in study D, the minimum reached mean sample BWC (%) value. Using equation (1) continuous severity values were calculated with severity = risk / likelihood. Further, risk levels for individual animals based on the humane endpoint threshold of -20% BWC (%) were calculated arithmetically as relative fractions.

3. RESULTS

Four studies were chosen for exemplary reasons (A-D). Although the focus of each study lies on different aspects, all were continuously monitoring body weight as an indicator of animal well-being on a daily basis. Study durations were different (t(A)=14 days, t(B)=6 days, t(C)=26 days and t(D)=116 days) and prospective severity was defined differently as well. Table 2 shows the number of total animals in the study along with the number of animals that had to be sacrificed or died due to treatment. In terms of risk assessment, the immediate proportion of animals that died is also an expression of risk for participating in the respective experiment. The table also lists the prospective severity each principal investigator had to define before study start. There is, however, a tangible gap between the number of deaths and the defined severity levels. Also, this assessment does not look at study duration or time at risk at all – neither did the original prospective assessment. Study A was considered to be "mild-moderate" in terms of severity and 9 animals out of 102 had to be sacrificed because of the mandatory -20% body weight threshold violation. Here, it remains unclear whether these animals had the potential to fully recover. Study D was considered "moderate". Here, zero out of 32 animals died within study duration of 116 days. This is remarkable because during this time 10 animals were below the -20% body weight change threshold for an extended period of time and were not sacrificed or did not die. There appears to be a notable qualitative difference in how treatment has an effect on animal well-being – it clearly shows that it is not generalizable.

All body weight data are longitudinal data sets, which are also called time-series. However, Figure 1 shows a time-free representation of just the data points. Body weight change values (%) are indexed for each study in which the order of appearance does not matter. With this representation it is easy to compare all four studies in terms of raw BWC state. Data points violating the -20% BWC threshold can easily be determined. Also, other thresholds can be tested as well. However, this approach is only reasonable if a threshold is the determinant for severity or any other defined humane endpoint. This strategy fails in case of study C where body weight is continuously increasing and animals have to be sacrificed early in the study. This kind of effect is not reflected at all in the parameter body weight change.

3.1 Study effects and distributions

The four studies in this paper were selected because they each represent a general case in severity assessment. As already mentioned, severity appears primarily to be study and therefore treatment dependent. This can easily be demonstrated with the carefully selected studies.

Study A shows a sharp drop in BW after treatment in comparison to the control group. Although study duration is not exceedingly long, a recovery can also be observed. Three states (normal, impaired, recovered) are easily discerned with the naked eye or simple statistics (Häger et al 2018).

Study B shows only small changes in BW but the animals are severely impaired in their well-being and die at relatively small changes in body weight. Treatment and control are hardly distinguishable by body weight alone.

Study C shows increasing body weight values during which animals died or had to be sacrificed. There are no indications from body weight alone that these animals are in some form of discomfort. On the contrary, if body weight alone was used to assess severity, this would have been misleading.

Study D shows an immediate but long-term and stable drop after treatment. Some animals even fall below the -20% BWC threshold but none of the animals died or had to be sacrificed during study duration. Animals were sacrificed after study termination so that it remains unclear whether they would have recovered. They did not die due to treatment.

This analysis is further backed-up by specifically analyzing the distribution of effects. For studies A to C the maximum reached mean of negative BWC for the dead animals was determined as well as the mean of the surviving animals. Each subsample was then resampled 1000 times in a bootstrapping process to determine the distribution of effect data. Both distributions were then plotted in Figure 3 to show hypothetical means (controls are green bars, dead animals are red bars). Since there are no dead animals in study D, the lowest reached mean BWC (%) of the treatment group was used instead.

Figure 3a shows that the means of control and largest effect groups are well separated. Both extrema are also well discriminated. However, transitions like BWC drop and recovery will move the red distribution back towards controls. This is a time-dependent effect as it was described elsewhere.

Figure 3b shows that animals are already dying at small changes in body weight. Although BWC reaches values as low as -26.34% half of the animals die in a BWC (%) range of [0;-15]. The published study (REF) reveals that animals show bad clinical scoring in this range. Animals are heavily impaired in their well-being and BWC reflects this only partially. There is also a small overlap in hypothetical distributions where it remains unclear to which group samples belong – nevertheless, exactly in this range animals are likely to die.

Figure 3c shows a large overlap in distributions in the hypothetical range [-5;0] BWC (%). In this study, body weight is increasing and animals experience severity rather early, however, at states with increased body weight. The sampling shows that, theoretically, the effect of treatment can reach lower BWC (%) values but in practice severity cannot be detected by body weight alone here. This body weight behavior is counter intuitive to most treatment effects where a drop in body weight is expected.

Figure 3d shows a long term drop in body weight even below the -20% BWC threshold, but no deaths. There is also some overlap between control and treatment group. For this study BWC (%) indicates severity but, obviously, the animals survive and are in no form of life-threatening discomfort. Here, BWC for severity assessment alone may even be wrong and would have led to unnecessary deaths.

In conclusion, assessing severity with body weight alone is futile. In some cases it is even impossible for body weight to act as a predictor for severity or even death. Body weight might be useful in some models but the way humane endpoints are determined may change (see Ref). It appears that not only different severity models are needed for specific study designs but also the role of body weight is much more complex than anticipated.

3.2 Density kernels and clustering

The time-free representation of study data in Figure 1 has the advantage of being rather intuitive at comparing body weight change values: since the data points oscillate around zero, kernel functions can be used to calculate density distributions. These functions smooth the actual data points to the given kernel (here: Gaussian) with a given bandwidth. The bandwidth is scaled so that it matches the standard deviation of the given kernel. Kernel smoothing can easily be achieved with the R function “density” from the R stats package. This approach is, however, surprisingly informative because it not only allows a pooling of different studies but also the calculation of likelihoods – a fact that leads to continuous risk assessment.

The representation in Figure 4a shows the density estimates of all four studies with a mean BWC (%) of zero. This can also be done for each study individually. But for showing the generalization power of clustering techniques the studies were pooled. Häger et al have shown that clustering can be a robust technique for finding individual severity levels. This can also be applied to the present data (here, only negative BWC (%) values were analyzed). A k-means algorithm with five clusters was applied to the pooled data resulting in five unequally sized clusters. The further away the clusters are from zero, the larger they become. This is an immediate reaction to smaller sample sizes in lower BWC regions. Cluster intervals are shown in Figure 4b: Cluster1[0; -1.05], Cluster2[-1.05; -2.69], Cluster3[-2.69; -5.71], Cluster4[-5.71; -14.35], Cluster4[-14.35; -34.37]. The number of clusters was not optimized as this was not needed here. As a result, the gaps between Clusters 1 to 3 are not large and a difference of 5% BWC is considered “acceptable” by NNN???. However, the clustering would be efficient with 3 clusters as well but the point is, that with this approach it is certainly possible to grade the individual severity of animals using BWC as the only parameter at any given time point into an arbitrary number of levels. Whether this reflects objective severity still remains open for debate as clustering can only reveal severity in context with other studies, where certain BWC levels are associated with effects and priorly determined prospective severities. In terms of risk towards the -20% BWC threshold the clusters certainly allow a clear discrimination of risk levels.

The k-means method is dependent on data distribution and it is eventually used to find agglomerations in the data. However, there are many other clustering techniques and the outcome of such a method must not necessarily be dependent on data alone. A very prominent application of determining individual levels in data is based on the golden ratio. It assumes a harmonic distribution of data based on the Fibonacci series and uses so called Fibonacci retracements (i.e., 100, 78.6, 61.8, 50, 38.2 and 23.6% etc) for data segmentation. The hypothesis behind this assumes that naturally occurring data follows the golden ratio. This technique is notoriously used as a quantitative method in the financial industry to determine price action. As stock prices are time-series as well, this method is easily transferred to assess severity levels in animal models between a given minimum and maximum of a parameter. Figure 4c shows the application of Fibonacci retracements on the kernel density representation of the four studies' body weight change data. In BWC (%) the retracement levels are R1 [0; -6.94], R2 [-6.94; -12.84], R3 [-12.84; -16.96], R4 [-16.96; -21.21], R5 [-21.21; -26.28]. Looking at the figure, it becomes clear that the retracements become much narrower the farther they are away from 100%. This has to do with the Fibonacci series whose fractions asymptotically approach 1. The main difference to the k-means approach is that the first retracement starts much later at BWC=-6.94% which allows some space for naturally occurring variance. Further, the levels are much more equally spaced and allow a finer grading of severity. Also, like in the stock market, the Fibonacci retracements appear to be following certain levels in the data. However, from a statistical point of view Fibonacci Retracements show no significant effect in beating the markets, therefore they are marked as being inefficient.

3.3 Continuous models and likelihood estimation

Density kernel data can further be used in building continuous probability models. For this, an empirical cumulative distribution function (ECDF) is applied to the density kernel. This will result in a sigmoidal likelihood curve of the data in which for each data point a probability of occurrence (likelihood) is given. Although it is possible to do this with pooled study data, each study is assessed individually. Additionally, data are divided into negative and positive BWC (%) groups. The ECDFs for both groups are then calculated and plotted in Figure 5. In order to avoid functions running into zero too soon, a threshold of 5 % BWC in either direction of zero BWC (%) was allowed. The resulting ECDFs show likelihoods for positive (black curve) and negative (red curve) groups in which the probabilities of positive values are called "not at risk" and the negative values "at risk". The model not only allows direct risk assessment but also the prediction of likelihoods for new data points (query points, blue x, Figure N). Figure 5a shows the ECDFs of study A. Both curves cross at approximately BWC=0 %. 5 % in either direction there is a gap in the ECDF in which it remains unclear whether a sample is at risk or not. This threshold is arbitrary and can be adapted. Although for some samples, there may be an uneven ratio in likelihoods tilting the odds to one specific direction. To demonstrate the power of this method, a common query point at BWC=-10% was tested in studies A, B and D. This results in a likelihood of P=0.92 for study A, P=0.78 for study B and P=0.63 for study D. From the small fractions at risk with the black curves it becomes clear that high BW values show very low likelihoods for being at risk. However, for study C there is a caveat. Since body weight values are increasing or do not move away from zero, the modeling of negative likelihoods is not feasible. There simply are no values in the negative BWC range. Finally, the calculated likelihoods can be used in a quantitative risk/severity assessment equation.

3.4 Individual and general risk assessment

For studies with deteriorating effects on body weight and a defined humane endpoint, risk calculation is rather simple. Figure 6 shows individual risk levels for each BWC data point with regard to the -20 % BWC threshold. Risk is simply the fraction of actual data value in regard to the threshold. As body weight goes down, risk goes up. 100% risk means that the animal has to be euthanized. 6a to c shows characteristic BWC curves for individual animals under treatment from studies A, B and D. After treatment BW drops and recovers over time (a and b). Figure 6 d shows a control with rather constant BW values around zero. Figure 6e shows the body weight course of an animal of study C. Body weight is increasing over time and the animal had to be euthanized on day 12. Classic risk assessment with the previously presented tools is not possible with these data. However, inverting the data allows some pseudo risk calculation in which the animal has a risk of 0.51. As this approach is not recommendable as a general method, further risk assessment was not followed for study C.

Threshold-based risk calculation is easy to achieve but does not reflect characteristic properties of a model such as value likelihood and the actual number of occurring deaths etc. These are better reflected by the bootstrapped sample distributions of the dead animals. The probability mass function of the bootstrapped values is a quantitative risk measurement, since only values of dead animals are sampled. At these values animals really die and the ecdf links this to a probability. In combination with the likelihood values from the continuous modeling it is possible to calculate risk and severity quantitatively using the classic risk formula (1).

Figure 7 shows (a) the likelihood of BWC value occurrence and (b) the bootstrapped risk distribution of dead animals in the studies A and B. For study D there are no dead animals, so instead the minimum mean treatment effect (BWC (%)) was used as maximum risk of the model. Otherwise, risk cannot be computed for this study. (c) At last, severity can be calculated for each of the three studies as a continuous parameter. For any given project this knowledge can be used for severity assessment. Principal investigators just need estimates for maximum loss of body weight due to treatment as well as the connected risk and/or severity. With this information and bootstrapping methods the missing variable can be modeled. Knowledge or estimates can come from already established models or hypotheses, very much the way as power calculations for study designs are done.

4. DISCUSSION

We think that there are at least three qualities in every single parameter that will yield first information about the actual risk and/or severity state of animals:

1. The general type or expected flow of the parameter in the study (up or down)
2. Data and model based classes of severity (i.e., BWC (%) clusters, Y-axis)
3. Transient severity phases (i.e., treatment effects and recovery, X-axis)

Looking carefully at these three points will already establish a good context on how the respective animal model will behave in terms of risk and severity assessment. If it is clear that animals will gain weight due to treatment, humane endpoint definition must be changed. Also, it can be important if data are spread out to a large range or cluster together at characteristic spots. Finally, it does matter if animals are able to recover or stay at prolonged levels of severity.

Although risk can be defined using equation (1) the exact nature of risk remains actually a matter of definition. All calculations will not free scientists from the definition of what is "bearable" for animals and what is not. However, looking at different models where animals die will give a basic understanding of what a single parameter can contribute as a predictor of death (risk or severity).

As we have shown for studies A and B, risk is determined with actual occurring deaths due to treatment. This can be quantified for study C as well but as body weight behaves counter intuitive it is unclear of how this is connected to severity numerically. In study D, however, risk is defined otherwise. Here, the lowest reached mean BWC (%) value was defined as highest risk. This can only remain some kind of expedient tool for severity characterization. If this was not followed, study D would show no risk and therefore no severity at all (since no animal died). Of course, this will look differently when regions of lower body weight are defined as risk factors. When this is done, the definition of severity changes as well: it too becomes transient in nature.

Another major issue of severity assessment comes with the comparison of different studies and time frames. As animals are at risk for different lengths of time it remains unclear how to handle this numerically. Surely it is somehow important how long animals suffer in a lower body weight region before giving the chance of recovery. In study D, 10 animals are well below the humane endpoint threshold for extended periods of time but they did not die – it remains unclear if and how much they suffer. Of course, short-term spikes in severity can have different effects on animal wellbeing (studies A to C) than long lasting ones (study D). From body weight alone this is hard to characterize if study outcomes show such diversity.

The exact definition of risk is therefore important (i.e., death or maximum/minimum values). As long as there are at least two values such as the actual data point and a threshold or maximum acceptable value the calculation of risk is easy. This become more difficult if the concept of risk is unclear or no exact threshold can be defined. In these cases the calculation of empirical or hypothetical likelihoods may help by comparing risk effects such as deaths from historical data and use them in prospective modeling, i.e. by using the bootstrapping method. Still, this would remain an educated guess but prospective power calculations for sample size determination suffer exactly from the same disadvantage when researchers have to guess an effect size.

In risk analysis there are mainly two approaches: firstly, potential future events are analyzed which may have a negative impact on animals or assets and secondly, making judgements on how much risk is tolerable when experimental factors are changed. Basically, risk analysis is the art of determining what can go wrong and what the consequences are. However, in animal experiments where the exact level of severity is hardly determinable or expressed as some form of subjective parameter, such as clinical scoring, this leaves scientists with the already mentioned dilemma. The only objective and countable consequence in animal experimentation is death which essentially reduces risk to a yes/no question as a function of severity. These problems are usually solved with generalized linear models such as logistic regression which is a legit way of modeling individual risk quantitatively. However, from a perspective of animal wellbeing this thinking is not entirely sufficient, because the animals may very well be in a state of extreme suffering where they are still alive and will not die. In these cases the regression model is not applicable and suffering cannot be measured unless scientists actively define a threshold for maximum bearable suffering.

In addition, in animal research severity is usually not expressed as a continuous variable or probability but as a category (i.e. low, medium and severe). As we have clearly shown with all four study examples, severity is a transient variable. Whenever severity is expressed as a category, the detection of transient states becomes impossible. This way, severity can only serve as a label for the highest achieved level of suffering in a given model which generalizes risk for all animals to a level that is not true for most of the participating animals of a study in either direction: suffering or recovery. A continuous model much better reflects these circumstances on an individual level. Therefore, it is desirable to describe risk not only as a yes/no question but also as a graded system for certain objective parameters. This will allow assessing how much animals are suffering at a given point in time. From this, actual risk can be calculated numerically.

A generally adopted framework for animal handling is the 3R-principle (refine, reduce and replace) which was introduced by Russel and Burch in 1959. Sometimes experimental effects become too severe or animals react negatively to exposure so they need to be euthanized to avert ongoing or prolonged suffering. A typical mandatory threshold, or humane endpoint, is the loss of 20% in body weight. Conditions near or below that threshold are considered to be severe while not regarding the actual state of the animal. Without further context this is highly dogmatic as animals may still have the potential for full recovery (see study D).

Severity is the quality of being unpleasant. In this definition the word quality is highly unclear and/or subjective as it is difficult to obtain objective responses from animals. There is also much variation, e.g. in behavioral parameters, indicating, that body weight is, at least to some extent, an ambiguous parameter. Basing decisions on animal deaths on one value alone may therefore be controversial. As it is difficult to define a measurement for suffering in animals - a deeper look into this topic may be valuable for getting a better understanding on how animals cope with impairing factors and how this information can be quantified and sorted objectively. A first approach was followed by Häger et al (2018) in which they correlated both, body weight and voluntary wheel running (VWR) performance revealing clusters of generalized severity. The model was built using a naive k-means clustering algorithm with multiple permutations for finding the optimal cluster borders for wheel running and the corresponding 95% confidence borders. Using this, other data can be assessed by projecting them into the model. A rather relevant feature of this approach is that for the first time (relative) severity levels can be attributed and new data can be assessed on an individual basis allowing model comparisons.

Clustering data allows individual segmentation of data. In terms of a specific hypothesis (e.g. data distribution or golden ratio etc.) this may even be called unbiased. However, finding severity ground truths remains a highly subjective matter. Finding severity levels appears to be possible only with heuristic methods and each study must be analyzed carefully in terms how animals are reacting to treatment. With this knowledge, finer risk and severity grading will become possible – even on an individual level.

But, for a n-parameter model, n parameters must also be measured, which is not always possible or desirable. The question that remains is: Does the future of severity assessment lay in complex multiple parameter models or can single parameters like body weight still tell a good story?

5. REFERENCES