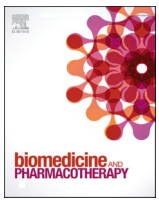
Biomedicine & Pharmacotherapy 160 (2023) 114317

Contents lists available at ScienceDirect

Biomedicine & Pharmacotherapy

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



Between therapy effect and false-positive result in animal experimentation

Paweł Sosnowski [a,1], Piotr Sass [a,1], Anna Stanisławska-Sachadyn [b], Michał Krzemiński [c], Paweł Sachadyn [a,*]

[a] Laboratory for Regenerative Biotechnology, Gdańsk University of Technology, ul. Narutowicza 11/12, 80-233 Gdańsk, Poland
[b] Department of Molecular Biotechnology and Microbiology, Gdańsk University of Technology, ul. Narutowicza 11/12, 80-233 Gdańsk, Poland
[c] Institute of Applied Mathematics, Faculty of Applied Physics and Mathematics, Gdańsk University of Technology, ul. Narutowicza 11/12, 80-233 Gdańsk, Poland

ARTICLE INFO

Keywords:
False-positive results
Sample size
Statistical significance
Animal experimentation, pharmacoregeneration
Machine-learning
Naïve Bayes classifier
Support Vector Machine classifier

ABSTRACT

Despite the animal models' complexity, researchers tend to reduce the number of animals in experiments for expenses and ethical concerns. This tendency makes the risk of false-positive results, as statistical significance, the primary criterion to validate findings, often fails if testing small samples. This study aims to highlight such risks using an example from experimental regenerative therapy and propose a machine-learning solution to validate treatment effects. The example analysed was the pharmacological treatment of ear pinna punch wound healing in mice. Wound closure data analysed included eight groups treated with an epigenetic inhibitor, zebularine, and eight control groups receiving vehicle alone, of six mice each. We confirmed the zebularine healing effect for all 64 pairwise comparisons between treatment and control groups but also determined minor yet statistically significant differences between control groups in five of 28 possible comparisons. The occurrences of significant differences between the control groups, regardless of standardised experimental conditions, indicate a risk of statistically significant effects in the case a compound lacking the desired biological activity is tested. Since the criterion of statistical significance itself can be confusing, we demonstrate a machine-learning algorithm trained on datasets representing treatment and control experiments as a helpful tool for validating treatment outcomes. We tested two machine-learning approaches, Naïve Bayes and Support Vector Machine classifiers. In contrast to the Mann-Whitney U-test, indicating enhanced healing effects for some control groups receiving saline alone, both machine-learning algorithms faultlessly assigned all animal groups receiving saline to the controls.

1. Introduction

Animal models are the primary tools used to develop novel therapies and evaluate their effectiveness. Selecting an appropriate sample size is critical to obtaining valid experimental evidence. I The ethical concerns expressed as the 3Rs principle (Replacement, Reduction, Refinement) [1], in addition to high expenses and extended time of observations, often tempt researchers to report results based on a single experiment involving a few animals. The advantage of this approach is that animal lives are saved, and discoveries are brought to light to be further verified by other scientists. Statistical tests are used to substantiate findings. However, the issue of statistically significant false positives is not always given due consideration. Reports on the low reproducibility of published data seem to signal the problem of false-positive results in biomedical research. For example, a review by Amgen confirmed 6 out of 53 (11%) pre-clinical trials [2]. Several statistical methods have been proposed to control the risk of false-positive results [3]. These statistical approaches though helpful, cannot nullify differences between experiment replicates resulting from technical and biological variations. The present study aims to point out the risks of false-positive results in animal experimentation on small samples and propose solutions to distinguish variation between experimental replicates from significant treatment effects.

To demonstrate the risk of overstating evidence from small samples, we discuss an example of experimental data we obtained in a previous study [4], where we proved that an epigenetic inhibitor, zebularine, induced ear pinna regeneration in mice. The principle of this model is that the regenerative response is evaluated based on measuring the closure of 2-mm holes made in the ear pinna. The closure was remarkable, 83.2 ± 9.4% in the zebularine-treated mice vs 43.7 ± 15.4% in the

* Corresponding author.
E-mail address: psach@pg.edu.pl (P. Sachadyn).
[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.biopha.2023.114317
Received 16 October 2022; Received in revised form 18 January 2023; Accepted 26 January 2023
Available online 1 February 2023

controls ($p < 10E-30$), and confirmed in several independent experiments. Unexpectedly, while revisiting the study results, we recorded moderate but statistically significant differences between control groups treated with the vehicle only. What is essential, our experiment was conducted in standardised conditions on a homogenous population of inbred mice of the same strain, sex, and age. Given that comparative statistical testing can produce confusing signals, we demonstrate machine-learning predictions using Naïve Bayes and support vector machine algorithms as effective tools for validating responses to drug treatment. We believe the presented analysis will provide valuable guidance on the design and interpretation of animal experiments.

2. Methods

2.1. Animal experiments

Ear punch experiments in mice are described in our previous publication [4]. The animal study protocols were approved by the Local Ethics Committee for Animal Experimentation (http://lke.utp.edu.pl) at the Faculty of Animal Breeding and Biology, University of Science and Technology, in Bydgoszcz, Poland (approval No. 5/2015). All experiments were performed in accordance with relevant guidelines and regulations.

2.2. Statistical analysis

Two-sample comparisons were carried out with the two-tailed non-parametric Mann-Whitney U-test. Power and sample size calculations were performed using the means and standard deviations determined for each tested animal group for the Mann-Whitney U-test. The sample sizes were calculated at a power of 0.8 and an alpha level of 0.05. The Kruskal-Wallis test with Dunn's post hoc analysis and Bonferroni correction were applied for multiple comparisons.

All statistical computations were done with XLSTAT (Addinsoft). The analysed research data representing the percentages of ear pinna hole closure are included in Supplemental Table S1.

2.3. Machine-learning experiments

The machine-learning experiment for determining statistical significance between randomly sampled groups was carried out using the R package [5] (https://www.R-project.org/). The two-sample Mann-Whitney U test was performed using the standard wilcox.test function, where the null hypothesis is that the two distributions of samples differ by a location shift 0. Successive tests in the experiments were conducted on replicated random samples sampled with replacement.

2.4. Naïve Bayes and support vector machine classification

The Naïve Bayes and Support Vector Machine Classifiers (XLSTAT, Addinsoft) were applied for assigning samples to control and treatment types. The classifiers were run with default parameters. For this prediction, ear hole closure areas in each sample were sorted in descending order, and the missing values were replaced with means (Supplemental Table S2). Sampled groups were obtained with Data sampling XLSTAT function using random sampling without replacement (Addinsoft).

3. Results

3.1. Significant differences between control groups determined by comparison tests

In the present study, we conducted a retrospective analysis of wound closure data from previous research [4] (Supplemental Table S1). The data represents eight zebularine-treated and eight control groups of six mice each (Fig. 1). First, using the Mann-Whitney test, we made two-sample comparisons between all study groups. Each group treated with zebularine showed significantly better wound closure than any control group (Fig. 2a). Unexpectedly, in 5 of 28 comparisons between the control groups treated with saline alone, we observed statistically significant differences. One of the p-values achieved $5.0 \times 10E-4$ (Fig. 2a). As the extent of healing was way lower in the control than in zebularine-treated mice, the pro-regenerative effect of zebularine was unquestionable (Fig. 2b). On the other hand, the incidents of significant differences between control groups receiving saline seem puzzling. As noted above, the significant differences between the control groups did not compromise the results found for the active compound zebularine. When observed in comparisons between saline-treated control groups, significant differences in regenerative effects are evident as false-positive results. However, the occurrence of such outcomes signals that statistically significant differences may be determined when testing a compound lacking the desired biological activity.

In order to include corrections for multiple comparisons, we conducted the Kruskal-Wallis test, followed by Dunn's pairwise tests. This analysis identified statistically significant differences between control groups in four of 28 pairwise comparisons (Fig. 2c). It is worth noting that the Kruskal-Wallis analyses were possible to perform after the completion of the experiments for several control groups. In experiments involving a single treatment and a single control group, as often practised for saving animals, multiple comparison tests cannot be applied.

3.2. Power statistics and sample size

Statistical significance is confronted with power and sample size statistics to validate study results. In our analysis, the statistical power

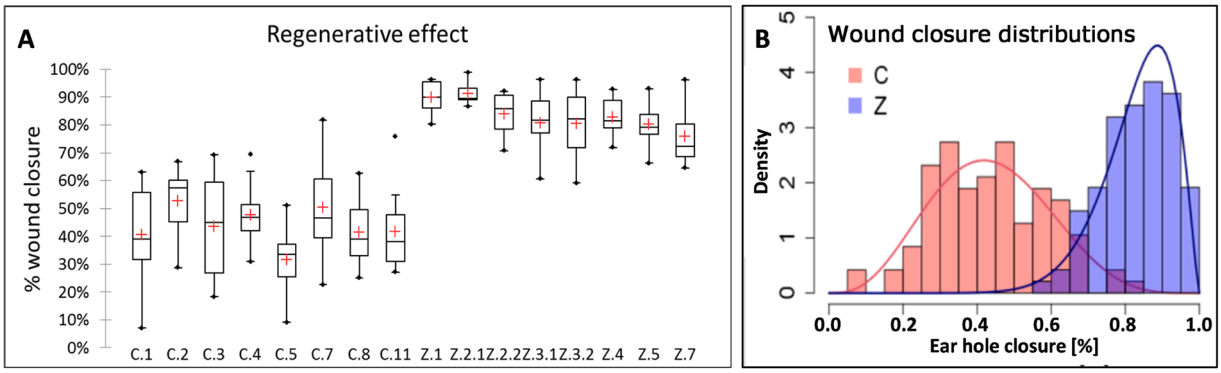


Fig. 1. Ear pinna wound healing - experiment results: A) the percentage of ear hole closure in zebularine (Z.1-Z.7) and saline-treated (C.1-C.8) mice. Each tested group consisted of six mice (n = 12 ears). The boxes represent the first and third quartiles; the bars indicate the minimum and maximum; the "+ " and "–"marks represent the means and medians, respectively; B) distributions of ear hole closure results in the controls (C, red) and zebularine-treated mice (Z, purple).

A) p-values for two-sample comparisons with the Mann-Whitney U-test

	C.1	C.2	C.3	C.4	C.5	C.7	C.8	C.11	Z.1	Z.2.1	Z.2.2	Z.3.1	Z.3.2	Z.4	Z.5	Z.7
C.1	1.0E+00															
C.2	1.6E-01	1.0E+00														
C.3	7.6E-01	2.4E-01	1.0E+00													
C.4	2.6E-01	3.1E-01	4.7E-01	1.0E+00												
C.5	2.7E-01	5.0E-04	1.3E-01	6.5E-03	1.0E+00											
C.7	2.2E-01	5.5E-01	4.4E-01	8.3E-01	1.0E-02	1.0E+00										
C.8	8.8E-01	3.7E-02	8.8E-01	2.5E-01	9.1E-02	2.6E-01	1.0E+00									
C.11	9.8E-01	2.4E-02	8.9E-01	1.2E-01	1.6E-01	1.8E-01	9.3E-01	1.0E+00								
Z.1	7.4E-07	7.4E-07	7.4E-07	3.6E-05	7.4E-07	1.5E-06	1.5E-06	7.4E-07	1.0E+00							
Z.2.1	7.4E-07	7.4E-07	7.4E-07	3.6E-05	7.4E-07	7.4E-07	1.5E-06	7.4E-07	7.6E-01	1.0E+00						
Z.2.2	7.4E-07	7.4E-07	7.4E-07	3.6E-05	7.4E-07	5.0E-05	1.5E-06	3.0E-06	6.0E-02	8.9E-02	1.0E+00					
Z.3.1	2.2E-05	3.7E-05	3.7E-05	2.2E-05	3.1E-06	3.0E-04	1.1E-05	1.2E-05	4.3E-02	2.1E-02	4.2E-01	1.0E+00				
Z.3.2	5.2E-06	8.9E-06	1.4E-05	1.4E-05	7.4E-07	2.7E-04	3.0E-06	8.9E-06	3.3E-02	2.0E-02	5.5E-01	9.4E-01	1.0E+00			
Z.4	7.4E-07	7.4E-07	7.4E-07	3.6E-05	7.4E-07	5.0E-05	1.5E-06	3.0E-06	1.7E-02	4.5E-03	8.4E-01	8.2E-01	8.0E-01	1.0E+00		
Z.5	7.4E-07	1.5E-06	1.5E-06	1.5E-06	7.4E-07	1.0E-04	1.5E-06	3.0E-06	2.3E-03	6.6E-04	4.4E-01	7.7E-01	7.1E-01	4.1E-01	1.0E+00	
Z.7	7.4E-07	1.5E-06	1.4E-05	8.9E-06	7.4E-07	2.3E-03	1.5E-06	5.0E-05	1.4E-03	1.1E-03	2.4E-02	2.3E-01	3.0E-01	3.9E-02	1.3E-01	1.0E+00

B) Size of the healing effect: differences in percentages of ear hole closure

	C.1	C.2	C.3	C.4	C.5	C.7	C.8	C.11	Z.1	Z.2.1	Z.2.2	Z.3.1	Z.3.2	Z.4	Z.5	Z.7
C.1	0.0															
C.2	12.1	0.0														
C.3	3.0	9.1	0.0													
C.4	7.0	5.0	4.0	0.0												
C.5	8.9	21.0	11.9	16.0	0.0											
C.7	9.8	2.3	6.8	2.7	18.7	0.0										
C.8	0.8	11.3	2.2	6.3	9.7	9.0	0.0									
C.11	1.0	11.1	2.0	6.0	9.9	8.8	0.2	0.0								
Z.1	49.4	37.3	46.4	42.4	58.3	39.6	48.6	48.4	0.0							
Z.2.1	50.6	38.5	47.6	43.6	59.5	40.9	49.8	49.6	1.2	0.0						
Z.2.2	43.4	31.4	40.4	36.4	52.3	33.7	42.7	42.4	6.0	7.2	0.0					
Z.3.1	40.1	28.0	37.1	33.1	49.0	30.3	39.3	39.1	9.3	10.5	3.3	0.0				
Z.3.2	40.0	28.0	37.0	33.0	48.9	30.3	39.3	39.0	9.4	10.6	3.4	0.1	0.0			
Z.4	42.3	30.2	39.3	35.3	51.2	32.6	41.5	41.3	7.1	8.3	1.1	2.2	2.3	0.0		
Z.5	39.8	27.7	36.8	32.8	48.7	30.0	39.0	38.8	9.6	10.8	3.6	0.3	0.2	2.5	0.0	
Z.7	35.4	23.3	32.4	28.3	44.3	25.6	34.6	34.4	14.0	15.3	8.1	4.7	4.7	7.0	4.4	0.0

C) p-values for multiple comparisons with the Kruskal-Wallis analysis and the post-hoc Dunn's-test

	C.1	C.2	C.3	C.4	C.5	C.7	C.8	C.11
C.1	1							
C.2	0.058	1						
C.3	0.756	0.113	1					
C.4	0.254	0.450	0.407	1				
C.5	0.147	**0.001**	0.078	**0.010**	1			
C.7	0.216	0.510	0.355	0.923	**0.007**	1		
C.8	0.992	0.062	0.754	0.261	0.159	0.223	1	
C.11	0.894	**0.042**	0.657	0.203	0.188	0.171	0.904	1

D) Power and sample size statistics for the control pairs showing significant differences

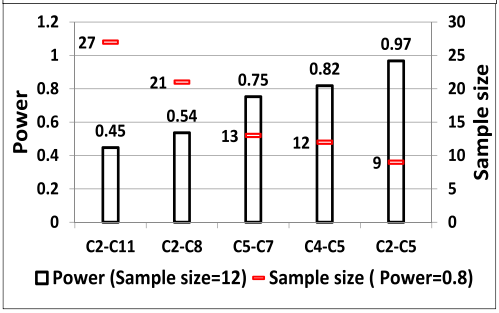


Fig. 2. Statistical significance determined in two-sample and multiple-sample comparisons contrasted with the effect size and the power and sample size statistics. A) Statistical significance: p-values computed for pairwise comparisons between experimental groups of zebularine-treated (Z.1-Z.7) and saline-receiving (C1-C.11) animals using the two-tailed Mann-Whitney test; the statistically significant differences between control groups (C1-C.11) are marked with red font. B) Size of the healing effect: the differences in mean percentages of ear hole closure between the experimental groups. Each tested group consisted of six mice (n = 12 wounds). C) Multiple comparisons between the control groups. The Kruskal-Wallis test, followed by Dunn's test for pairwise comparisons, was performed for eight control groups of mice receiving the vehicle alone. P-values below 0.05 are highlighted with a red background. The Bonferroni corrected significance level was set at 0.0018; the p-value significant after the correction was emphasised with bold font. D) Statistical power and sample sizes calculated for the pairwise comparisons between the control groups of mice receiving the vehicle alone and showing significant differences. The sample sizes were calculated at a power of 0.8 and an alpha level of 0.05; the statistical power values were determined for the sample size n = 12 (12 ear wounds from six mice). The means and standard deviations determined for each tested group were used for the computation.

values calculated for the comparisons showing significant differences between the control groups were as high as 0.82 and 0.97 (Fig. 2d). Also, the sample size in our experiments proved sufficient for the differences we observed. The sample sizes to achieve a power value of 0.8 were n = 9 (C2 vs C5 group) and n = 12 (C4 vs C5 group) (Fig. 2d), which was equal to or less than the sample size we applied in the experiment (n = 12 ears).

3.3. Machine-learning experiment

Assuming that all collected control samples are the estimate of non-drug-induced ear pinna wound closure in our experimental setup, we designed a machine-learning experiment to model the risk that two randomly sampled groups show significant differences. The control-control experiment consisted in randomly sampling 1000 k-element samples from all control observations, with k ranging from 3 to 12. Each of the 1000 samples was then compared using the Mann-Whitney U-test with 1000 other samples of the same size randomly sampled from all control observations. Then mean numbers of insignificant results per million were computed at $p = 0.05$ and $p = 0.01$ for each k. Next, we carried out an analogous experiment for comparisons between randomly sampled treatment and control groups, where each k-sample from the treatment group was tested against 1000 samples from the control group of the same size.

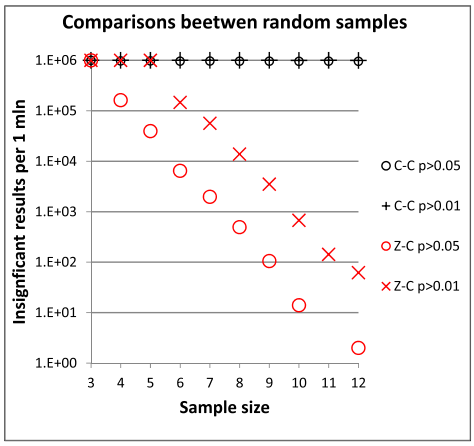
The machine-learning experiments demonstrated that the frequencies of false-negative results (insignificant treatment effect compared to controls) and false-positive results (significant differences between controls) were shallow, provided the sample size was adequate (Fig. 3). While the incidence of false-negative results markedly decreased with increasing sample size, the frequencies of false-positives were roughly similar for sample sizes k = 4–12, ranging from 2.7% to 4.6% (at $p = 0.05$, Fig. 3). Interestingly, for 3-element samples, the frequency of significant results between control groups equalled zero; however, the same was the chance for significant differences between control and treatment groups.

3.4. Variation between treatment and control groups

Fig. 1a shows the inter-group variation. While there is a clear difference in wound closure effect of (83.2% vs 43.7%) between the assembled treatments and controls, the differences between all pairs of treatment and control groups range from 23.3% to 59.5%, and those between all pairs of control groups from 0.20% to 21.0% (Fig. 2b). Consequently, it is challenging to delimitate a satisfactory difference between a single control group and a single treatment group. Overlapping distributions of control and treatment observations illustrate this problem (Fig. 1b).

3.5. Naïve Bayes classification

The main principle of applying the Naïve Bayes method is to use the prior probability obtained for training data to compute the posterior



Insignificant results per 1 mln comparisons

Sample size	C-C p>0.05	C-C p>0.01	Z-C p>0.05	Z-C p>0.01
3	1 000 000	1 000 000	1 000 000	1 000 000
4	972 838	1 000 000	161 703	1 000 000
5	968 867	1 000 000	39 474	999 993
6	959 800	996 135	6 435	145 899
7	962 483	995 778	1 977	56 365
8	959 397	994 079	492	13 842
9	958 574	993 865	105	3 508
10	955 826	993 713	14	673
11	960 181	994 655	0	142
12	953 943	992 945	2	62

Fig. 3. Machine-learning experiment for the assessment of insignificant comparisons made using the Mann-Whitney U-test. The comparisons were performed for 1 mln pairs sampled randomly from controls (C-C) and 1 mln pairs randomly sampled from control and treatment groups (C-Z). Low frequencies of insignificant control-treatment comparisons describe a low risk of false-negative results. High frequencies of insignificant inter-control comparisons depict a low risk of false-positive results.

probability. In our setting, the differences in the distributions of the wound closure between the control and treatment groups allow discrimination of subsequent test samples. In other words, the Bayesian approach was used here to learn the difference between the outcomes of effective therapeutic interventions and control experiments.

We applied a Gaussian Naïve Bayes Classifier from the XLSTAT package [6,7]. In our experimental setup, the observations obtained for control and treatment groups create a training dataset. The training dataset constitutes a prior for learning an algorithm that assigns any subsequently tested group into either a control or treatment type. We performed three experiments using the Naïve Bayes Classifier.

In the first experiment, we generated 1000 12-element groups randomly sampled from all control observations (Supplemental Table S3) and 1000 12-element groups randomly sampled from all treatment observations (Supplemental Table S4). Then we sorted out the sampled control group achieving the best mean wound closure and the sampled treatment group showing the worst mean wound closure (Supplemental Table S2). Next, we classified these two sampled groups (prediction set) into control or treatment types using a supervised machine learning algorithm based on the training set of eight control and eight zebularine-treated samples (Table 1). As determined using the Mann-Whitney U-test, the randomly sampled control group displaying the best wound closure showed significant differences relative to 5 of 8 experimental control groups (i.e. the sets of observations recorded during experiments as listed in Table S1) and the assembly of all controls. At the same time, the Bayes algorithm classified the sample into a control type. An analogous analysis was carried out for the randomly sampled treatment group with the lowest mean wound closure. The Mann-Whitney U-test revealed significant differences in 4 of 8 comparisons to experimental treatment groups and the assembly of all treatment observations, while the Bayes algorithm classified this group into a treatment type (Table 1). Fig. 4 shows a graphical representation of the training and the prediction sets described above for an intuitive explanation of Bayes classification.

Table 1

Naïve Bayes and SVM classifier predictions for the best-closing sampled control and worst-closing sampled treatment group contrasted with the results of the Mann-Whitney U-test. The best closing control and the worst closing treatment groups were sorted out each from 1000 12-element groups randomly sampled from assembled controls and assembled treatments, respectively. As determined with the Mann-Whitney U-test, the best closing sampled control displayed significant differences compared to C.1, C.3, C.5, C.8, C.11 and assembled controls, while classified into a control type in the Bayes prediction based on the training set, including eight experimental control samples and eight experimental treatment samples. Analogously, the worst closing sampled treatment group was classified into a treatment type in the Bayes prediction, though it showed significant differences in two-sample comparisons with Z.1, Z.2.1, Z.2.2., Z.4, and assembled treatments.

Sample	% closure	Mann-Whitney U-test		Machine-learning classifier		
		p-val vs best closing sampled control	*p*-val vs worst closing sampled treatment	Dataset	Naïve Bayes	Support Vector Machine (SVM)
Best closing sampled control	57.7%	1.00E+ 00	2.74E-04	Prediction	**Control**	**Control**
Worst closing sampled treatment	75.7%	2.74E-04	1.00E+ 00	Prediction	**Treatment**	**Treatment**
Zebularine 200 mg/kg	74.2%	1.44E-04	6.30E-01	Prediction	**Treatment**	**Treatment**
Zebularine 500 mg/kg	79.6%	1.41E-05	3.19E-01	Prediction	**Treatment**	**Treatment**
C.1	40.5%	**1.65E-02**	7.40E-07	Training	Control	Control
C.2	52.6%	3.54E-01	1.48E-06	Training	Control	Control
C.3	43.5%	**4.62E-02**	8.88E-06	Training	Control	Control
C.4	47.6%	6.77E-02	5.18E-06	Training	Control	Control
C.5	31.6%	**2.44E-05**	7.40E-07	Training	Control	Control
C.7	50.3%	2.47E-01	1.43E-03	Training	Control	Control
C.8	41.3%	**5.62E-03**	1.48E-06	Training	Control	Control
C.11	41.5%	**3.85E-03**	3.33E-05	Training	Control	Control
Assembled controls	43.7%	**2.82E-03**	1.25E-07			
Z.1	89.9%	7.40E-07	**1.58E-04**	Training	Treatment	Treatment
Z.2.1	91.1%	7.40E-07	**5.18E-06**	Training	Treatment	Treatment
Z.2.2	84.0%	2.96E-06	**2.31E-02**	Training	Treatment	Treatment
Z.3.1	80.6%	2.07E-04	2.28E-01	Training	Treatment	Treatment
Z.3.2	80.6%	2.74E-04	1.83E-01	Training	Treatment	Treatment
Z.4	82.9%	2.96E-06	**3.19E-02**	Training	Treatment	Treatment
Z.5	80.3%	8.88E-06	2.03E-01	Training	Treatment	Treatment
Z.7	75.9%	6.56E-04	9.47E-01	Training	Treatment	Treatment
Assembled treatments	83.2%	2.34E-07	**7.42E-03**			

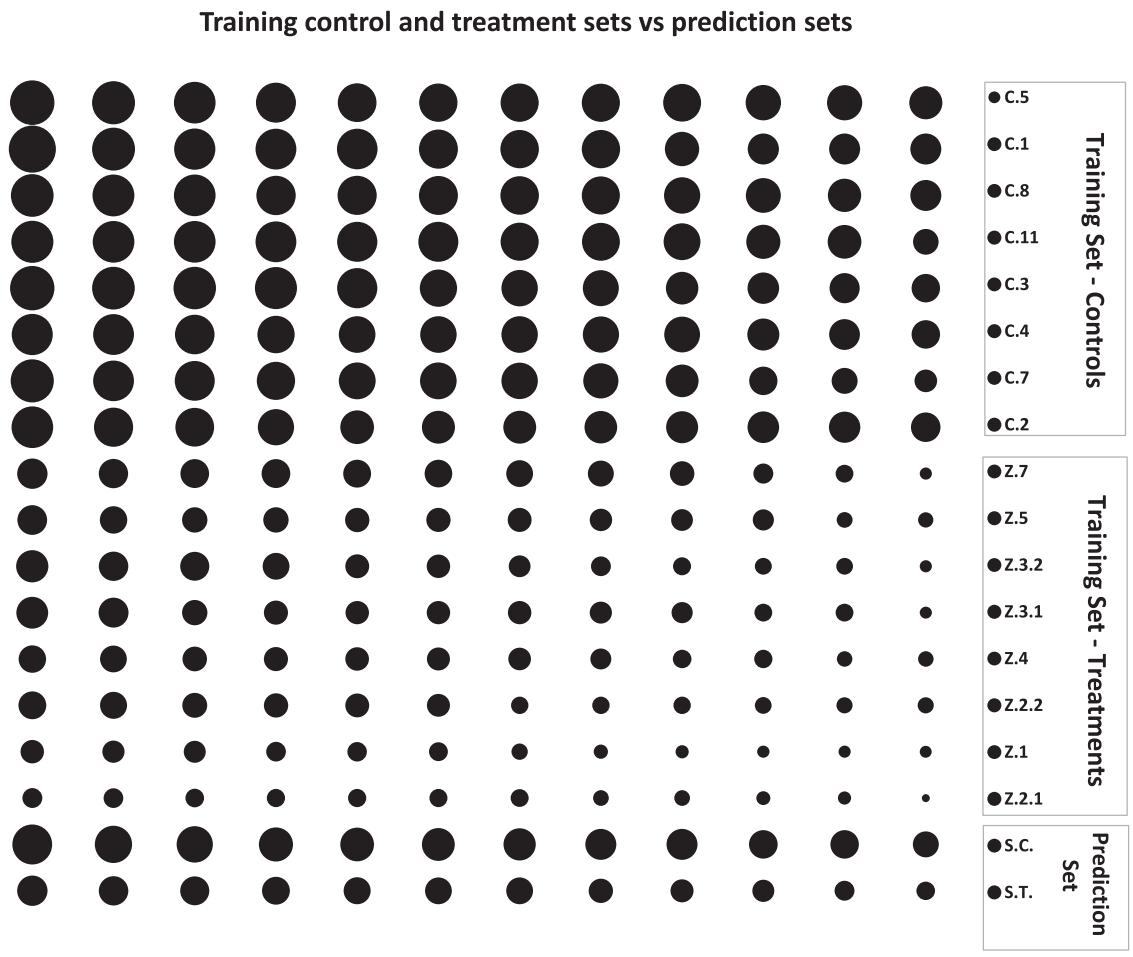


Fig. 4. Graphical depiction of the training dataset contrasted with the prediction dataset for Naïve Bayes and SVM classification. The circles represent ear pinna wounds, and the size of each circle corresponds to the area of each wound. S.C. – sampled controls showing maximal closure, S. T. sampled treatments showing minimal closure.

Table 2
Naïve Bayes and SVM classifier predictions for the best-closing experimental control (C.2) and worst-closing experimental treatment group (Z.7) contrasted with the results of the Mann-Whitney U-test. As determined with the Mann-Whitney U-test, C.2 showed significant differences compared with C.5, C.8, C.11, and assembled controls while classified into control types in the Bayes prediction based on the training set encompassing seven remaining control samples and seven remaining treatment samples. Analogously, Z.7 was classified as a treatment type in the Bayes prediction though it displayed significant differences in two-sample comparisons with Z.1, Z.2.1, Z.2.2, Z.4, and assembled treatments.

Sample	% closure	Mann-Whitney U-test		Machine-learning classifier prediction		
		p-val vs C.2	p-val vs Z.7	Dataset	Naïve Bayes	Support Vector Machine (SVM)
C.2	52.6%	1.00E+ 00	1.48E-06	**Prediction**	**Control**	**Control**
Z.7	75.9%	1.48E-06	1.00E+ 00	**Prediction**	**Treatment**	**Treatment**
C.1	40.5%	1.60E-01	7.40E-07	Training	Training	Training
C.3	43.5%	2.42E-01	1.41E-05	Training	Training	Training
C.4	47.6%	3.11E-01	8.88E-06	Training	Training	Training
C.5	31.6%	**4.96E-04**	7.40E-07	Training	Training	Training
C.7	50.3%	5.51E-01	2.32E-03	Training	Training	Training
C.8	41.3%	**3.74E-02**	1.48E-06	Training	Training	Training
C.11	41.5%	**2.42E-02**	4.96E-05	Training	Training	Training
Assembled controls	42.4%	**2.20E-02**	2.64E-07			
Z.1	89.9%	7.40E-07	**1.43E-03**	Training	Training	Training
Z.2.1	91.1%	7.40E-07	**1.11E-03**	Training	Training	Training
Z.2.2	84.0%	7.40E-07	**2.42E-02**	Training	Training	Training
Z.3.1	80.6%	3.71E-05	2.28E-01	Training	Training	Training
Z.3.2	80.6%	8.88E-06	2.98E-01	Training	Training	Training
Z.4	82.9%	7.40E-07	**3.87E-02**	Training	Training	Training
Z.5	80.3%	1.48E-06	1.28E-01	Training	Training	Training
Assembled treatments	84.3%	4.58E-08	**5.32E-03**			

We used the same training set in the second experiment to classify two other experimental groups. These groups received reduced zebularine doses of 500 and 200 mg/kg, while the groups in the training set were administered 1000 mg/kg. Both groups receiving the reduced doses were classified into a treatment type though 200 mg/kg showed a markedly lower wound closure effect than observed in the remaining treatment groups (Table 1).

In the third experiment, the Bayesian predictions were made for the experimental groups singled out from the learning set; the controls with the best and the treatments with the worst wound closure (C.2, Z.7). The training set included seven remaining experimental control and seven remaining experimental zebularine-treated groups. Though showing significant differences in two-sample comparisons with other control and treatment groups, the experimental control group with the best mean wound closure (C.2) and the experimental treatment group with the worst mean wound closure (Z.7) were assigned to control and treatment types, respectively (Table 2). In addition, we performed cross-validation, i.e. a series of comparisons where each experimental group was singled out from the training set and confronted with thus reduced training set containing the remaining control and treatment groups. As expected, each experimental control group was classified into a control type and each experimental treatment group into a treatment type.

3.6. Support vector machine classification

Analogous prediction tests as those made with the Naïve Bayes Classifier, were performed with a support vector machine (SVM) algorithm. The modern version of the SVM method was developed in the mid-90 s [8,9]. SVM belonging to the most popular and most robust supervised learning algorithms used for classification has found multiple applications in various fields, including medicine, biology and biotechnology [10]. The same training datasets were used, and identical prediction results were obtained using the Naïve Bayes and SVM classifiers (Tables 1 and 2). Cross-validation conducted as described in the previous section classified all experimental treatment groups into a treatment type and all control groups into a control type.

3.7. Machine-learning predictions for reduced training sets

Multiple replicates are not desirable owing to the requirement to decrease the number of animals in experiments. Obtaining large training datasets for machine-learning predictions is not always feasible. Therefore, we tested the Naïve Bayes and SVM classifiers using drastically reduced training datasets. We analysed two variants reflecting the extreme challenges. In the first one, we constructed a training dataset consisting of three control groups achieving the best mean ear hole closure (C.2, C.4, C.7) and three treatment groups achieving the worst mean ear hole closure (Z.3.2, Z.5, Z.7), as depicted in Fig. 5. The remaining control and treatment groups entered the prediction set. While statistically significant differences between the assembled controls of the training dataset and the control groups from the prediction set were determined using two-sample comparisons with the Mann-Whitney U-test, the Naïve Bayes and SVM classifiers properly assigned all groups to treatment and control types (Table 3). As the second variant, we performed an analogous analysis for the training set consisting of three control groups achieving the worst mean ear hole closure (C.1, C.5, C.8) and three treatment groups achieving the best mean ear hole closure (Z.1, Z.2.1, Z.2.2), as indicated in Fig. 5. Also in this variant of analysis, the Naïve Bayes and SVM classifiers correctly and consistently assigned all groups from the prediction set to control and treatment types (Table 4).

4. Discussion

In principle, a statistically significant result of comparison between treatment and control samples is the primary criterion accepted to substantiate findings. Our wound healing data analysis revealed statistically significant differences between independent replicates representing control groups of mice treated with vehicle alone. Moreover, we determined that the power values were satisfactory and sample sizes were adequate for the tests. The variations between the control groups we observed may seem unexpected since the animals of the same sex, age, and strain were maintained in standardised conditions and were

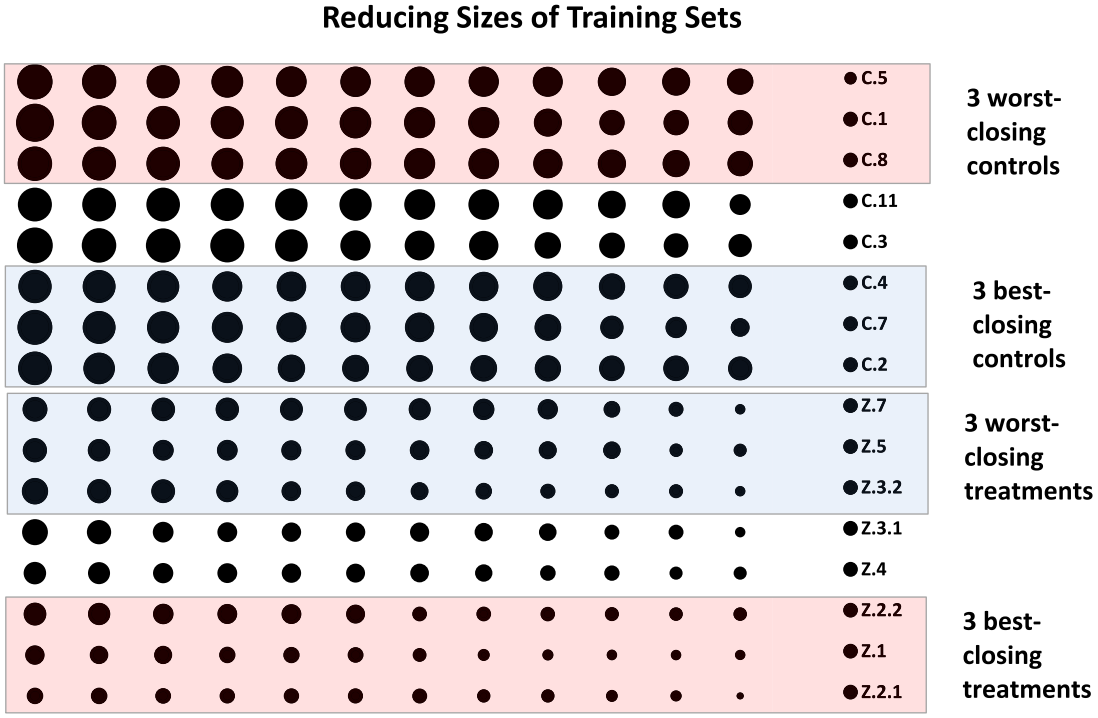


Fig. 5. Graphical depiction of the training set reduction for Naïve Bayes and SVM classification. The circles represent ear pinna wounds, and each circle's size corresponds to the wound's area; the blue and pink boxes indicate two variants of training datasets.

Table 3
Naïve Bayes and SVM classifier predictions for reduced training sets consisting of three control groups achieving the best mean ear hole closure (C.2, C.4, C.7) and three treatment groups achieving the worst mean ear hole closure (Z.3.2, Z.5, Z.7) contrasted with the results of the Mann-Whitney U-test.

Sample	% closure	Mann-Whitney U-test		Machine-learning classification		
		vs assembly of 3 best-closing control groups C.2, C.4, C.7	vs assembly of 3 worst-closing treatment groups Z.3.2, Z.5, Z.7	Dataset	Naive Bayes	Support Vector Machine
C.2	50.2%	1.00E+ 00	3.67E-11	Training	Control	Control
C.4				Training	Control	Control
C.7				Training	Control	Control
Z.3.2	78.9%	3.67E-11	1.00E+ 00	Training	Treatment	Treatment
Z.5				Training	Treatment	Treatment
Z.7				Training	Treatment	Treatment
Worst closing sampled Zeb 1000	57.7%	4.57E-06	2.95E-01	Prediction	Treatment	Treatment
Best closing sampled Ctrl 1000	75.7%	1.00E-01	1.21E-06	Prediction	Control	Control
Zebularine 200 mg/kg	74.2%	4.57E-06	1.34E-01	Prediction	Treatment	Treatment
Zebularine 500 mg/kg	79.6%	1.60E-06	8.40E-01	Prediction	Treatment	Treatment
C.1	40.5%	1.13E-01	2.01E-10	Prediction	Control	Control
C.3	43.5%	2.58E-01	5.57E-09	Prediction	Control	Control
C.5	31.6%	**2.33E-04**	2.88E-07	Prediction	Control	Control
C.8	41.3%	**5.84E-02**	5.74E-11	Prediction	Control	Control
C.11	41.5%	**3.31E-02**	1.44E-08	Prediction	Control	Control
Z.1	89.9%	3.27E-07	8.22E-04	Prediction	Treatment	Treatment
Z.2.1	91.1%	2.88E-07	3.71E-04	Prediction	Treatment	Treatment
Z.2.2	84.0%	7.79E-07	1.31E-01	Prediction	Treatment	Treatment
Z.3.1	80.6%	2.57E-06	4.75E-01	Prediction	Treatment	Treatment
Z.4	82.9%	7.79E-07	1.86E-01	Prediction	Treatment	Treatment

Table 4
Naïve Bayes and SVM classifier predictions for reduced training sets consisting of three control groups achieving the worst mean ear hole closure (C.1, C.5, C.8) and three treatment groups achieving the best mean ear hole closure (Z.1, Z.2.1, Z.2.2) contrasted with the results of the Mann-Whitney U-test.

Sample	% closure	Mann-Whitney U-test		Machine-learning classification		
		vs assembly of 3 worst-closing control groups C.1, C.5, C.8	vs assembly of 3 best-closing treatment groups Z.1, Z.2.1, Z.2.2	Dataset	Naive Bayes	Support Vector Machine
C.1				Training	Control	Control
C.5	37.8%	1.00E+ 00	0.00E+ 00	Training	Control	Control
C.8				Training	Control	Control
Z.1				Training	Treatment	Treatment
Z.2.1	88.3%	0.00E+ 00	1.00E+ 00	Training	Treatment	Treatment
Z.2.2				Training	Treatment	Treatment
Worst closing sampled Zeb 1000	57.7%	2.87E-11	1.32E-05	Prediction	Treatment	Treatment
Best closing sampled Ctrl 1000	75.7%	**2.24E-04**	1.15E-10	Prediction	Control	Control
Zebularine 200 mg/kg	74.2%	2.87E-11	3.01E-07	Prediction	Treatment	Treatment
Zebularine 500 mg/kg	79.6%	2.87E-11	7.03E-04	Prediction	Treatment	Treatment
C.2	52.6%	**3.78E-03**	2.87E-11	Prediction	Control	Control
C.3	43.5%	3.90E-01	2.87E-11	Prediction	Control	Control
C.4	47.6%	**3.72E-02**	2.88E-07	Prediction	Control	Control
C.7	50.3%	**3.83E-02**	2.78E-09	Prediction	Control	Control
C.11	41.5%	5.81E-01	1.15E-10	Prediction	Control	Control
Z.3.1	80.6%	4.76E-07	1.47E-02	Prediction	Treatment	Treatment
Z.3.2	80.6%	3.44E-10	3.60E-02	Prediction	Treatment	Treatment
Z.4	82.9%	2.87E-11	2.62E-02	Prediction	Treatment	Treatment
Z.5	80.3%	2.87E-11	3.19E-03	Prediction	Treatment	Treatment
Z.7	75.9%	2.87E-11	2.66E-04	Prediction	Treatment	Treatment

randomly distributed into control and treatment groups prior to the experiments. In addition, the model of ear punching we used is a straightforward procedure. Still, the results were not misleading because the effects recorded in the treatment groups receiving the tested agent, zebularine, were dramatically higher compared to the controls.

Nevertheless, our example demonstrates a potential risk of false-positive results in animal experimentation. Suppose an agent lacking the desired biological activity is tested; then a statistically significant effect can only manifest variation between experiments, likely due to the animal models' complexity. False-positive results, known as type I errors, are not uncommon in experimental research [11]. Different causes, including small sample sizes, increase the risk of false-positive findings [12–15]. As the measurements based on the determination of ear hole area using image analysis we carried out were reproducible, biological variation among the tested animals appears to be more likely to produce the observed differences between the control groups of animals. It is important to note that the differences occurred in standardised conditions, even though the mice were inbred and of the same age and sex. However, even inbred mice are known to display variation resulting from epigenetic divergence and random mutations [16,17]. Differences can be recorded, e.g., in gene methylation and expression levels between individual animals [18–20].

The presented example shows that standard tests to examine statistical significance and power and sample size statistics may be

insufficient to discriminate a response to treatment from a variation between controls. We point out, however, that the response was remarkably lower in the control animals than in the treatment groups. The differences in ear hole closure between the zebularine-treated and control groups ranged from 23.3% to 59.5% (mean 39.5%), while those between the control groups from 0.20% to 21.0% (mean 7.2%) (Fig. 2b). The evaluation is evident if the effect of the tested compound markedly exceeds the maximum difference recorded between the control groups. However, it may be challenging to determine the necessary margin discriminating between the treatment effect and biological or experimental variation. In such a case, a Bayes solution can be helpful.

In his essay from 1763, Thomas Bayes proposed that the probability of a future event can be determined based on past events. The concept was formulated as Bayes's theorem by Laplace in 1825. In the Bayesian approach, previously collected is used as a prior for analysing subsequent observations. Bayesian analysis has been successfully applied across different fields, including biomedical research [21]. In our study, the Bayesian approach involved learning an algorithm trained on several datasets from treatment and control experiments to assign the subsequent experiments to either treatment or control types. Although a new experimental result can be related to previously collected data using standard statistical tests for sample comparisons, the Bayesian approach has an essential advantage. Standard statistical tests for sample comparison allow computing statistical significance based on pre-defined distributions, thus creating more or less rigorous criteria to discriminate between positive and negative results. In Bayesian statistics, the background data are expressed as distributions to evaluate new observations; thus, the criteria adjust to what is learned from previous observations. In studies on animal models, such learning from previously collected data may save animals' lives.

The Gaussian Naïve Bayes classifier we applied is a robust method which does not require much training data. We showed that this tool effectively distinguished the response to treatment from variation between the control experiments. Such a practice may save time and expenses, especially in preliminary testing.

The Bayes predictions were confirmed with another machine-learning approach, support vector machine (SVM). We demonstrated that the SVM and Naïve Bayes Classifier equally evaluated the observed treatment effects. It is worth noting, however, that machine-learning classifiers are sensitive to the parameters applied, such as the type of distribution in the case of Naïve Bayes and the type of kernel in the case of SVM.

5. Conclusions

Using a spectacular example from animal experimentation, we highlight that a result of a single experiment on a small sample, although statistically significant, may not be biologically important. However, in principle, we would not like to question reporting of small sample experiments, especially if they are part of multi-faceted evidence. The leading conclusion of our study is that accounting for the variation between the control groups in a given experimental model helps proper evaluation of treatment effects, and machine-learning methods like Naïve Bayesian and Support Vector Machine classifiers can successfully solve this task.

Ethics approval

The animal experiments were approved by the Local Ethics Committee for Animal Experimentation in Bydgoszcz, Poland (approval No. 5/2015).

CRediT authorship contribution statement

PiSa and PaSo - data acquisition, discussion of results, and critical revision of the manuscript; AS-S -assistance with the design of the analysis, statistical consultancy and critical revision of the manuscript; M. K. - machine-learning experiment using the R-package, statistical consultancy and critical revision of the manuscript; PaSa - the concept, data analysis, experiment using a Naïve Bayes and Support Vector Machine classifiers, manuscript drafting and critical revision of the manuscript.

Conflict of interest statement

The authors declare that there are no conflicts of interest.

Acknowledgements

This study was supported by the grant of the National Centre for Research and Development of Poland "BIONANOVA", No. TECHMATSTRATEG2/410747/11/NCBR/2019. The funding source had no role in the writing of the manuscript or in the decision to submit it for publication.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.biopha.2023.114317.

References

[1] J. MacArthur Clark, The 3Rs in research: a contemporary approach to replacement, reduction and refinement, Br. J. Nutr. 120 (s1) (2018) S1–S7.
[2] C.G. Begley, L.M. Ellis, Drug development: raise standards for preclinical cancer research, Nature 483 (7391) (2012) 531–533.
[3] D. Colquhoun, The false positive risk: a proposal concerning what to do about p-values, Am. Stat. 73 (sup1) (2019) 192–201.
[4] P. Sass, P. Sosnowski, J. Podolak-Popinigis, B. Górnikiewicz, J. Kamińska, M. Deptuła, E. Nowicka, A. Wardowska, J. Ruczyński, P. Rekowski, P. Rogujski, N. Filipowicz, A. Mieczkowska, G. Peszyńska-Sularz, Ł. Janus, P. Skowron, A. Czupryn, P. Mucha, A. Piotrowski, S. Rodziewicz-Motowidło, M. Pikuła, P. Sachadyn, Epigenetic inhibitor zebularine activates ear pinna wound closure in the mouse, EBioMedicine 46 (2019) 317–329.
[5] R.C. Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, 2013. ⟨http://www.R-project.org/⟩.
[6] H. Zhang, The optimality of naive Bayes, Aa 1 (2) (2004) 3.
[7] C. Gertz, A. Gertz, B. Matthäus, I. Willenberg, A systematic chemometric approach to identify the geographical origin of olive oils, Eur. J. Lipid Sci. Technol. 121 (12) (2019), 1900281.
[8] B.E. Boser, I.M. Guyon, V.N. Vapnik, A training algorithm for optimal margin classifiers, Proceedings of the fifth annual workshop on Computational learning theory, 1992, pp. 144–152.
[9] C. Cortes, V. Vapnik, Support-vector networks, Mach. Learn. 20 (3) (1995) 273–297.
[10] J. Cervantes, F. Garcia-Lamont, L. Rodríguez-Mazahua, A. Lopez, A comprehensive survey on support vector machine classification: applications, challenges and trends, Neurocomputing 408 (2020) 189–215.
[11] R. Fricker, False positives are statistically inevitable, Science 351 (6273) (2016) 569–570.
[12] W. Forstmeier, E.J. Wagenmakers, T.H. Parker, Detecting and avoiding likely false-positive findings–a practical guide, Biol. Rev. 92 (4) (2017) 1941–1968.
[13] H. Ledford, Animal studies produce many false positives, Nature (2013).
[14] M. Macleod, Why animal research needs to improve, Nature 477 (7366) (2011), 511-511.
[15] E.S. Sena, H.B. Van Der Worp, P.M. Bath, D.W. Howells, M.R. Macleod, Publication bias in reports of animal stroke studies leads to major overstatement of efficacy, PLoS Biol. 8 (3) (2010), e1000344.
[16] H. Oey, L. Isbel, P. Hickey, B. Ebaid, E. Whitelaw, Genetic and epigenetic variation among inbred mouse littermates: identification of inter-individual differentially methylated regions, Epigenetics Chromatin 8 (1) (2015) 54.
[17] C. Weinhouse, O.S. Anderson, T.R. Jones, J. Kim, S.A. Liberman, M.S. Nahar, L. S. Rozek, R.L. Jirtle, D.C. Dolinoy, An expression microarray approach for the identification of metastable epialleles in the mouse genome, Epigenetics 6 (9) (2011) 1105–1113.
[18] B. Górnikiewicz, A. Ronowicz, M. Krzemiński, P. Sachadyn, Changes in gene methylation patterns in neonatal murine hearts: implications for the regenerative potential, BMC Genom. 17 (1) (2016) 231.

[19] J. Podolak-Popinigis, A. Ronowicz, M. Dmochowska, A. Jakubiak, P. Sachadyn, The methylome and transcriptome of fetal skin: implications for scarless healing, Epigenomics 8 (10) (2016) 1331–1345.
[20] B. Górnikiewicz, A. Ronowicz, P. Madanecki, P. Sachadyn, Genome-wide DNA methylation profiling of the regenerative MRL/MpJ mouse and two normal strains, Epigenomics 9 (8) (2017) 1105–1122.
[21] R. van de Schoot, S. Depaoli, R. King, B. Kramer, K. Märtens, M.G. Tadesse, M. Vannucci, A. Gelman, D. Veen, J. Willemsen, Bayesian statistics and modelling, Nat. Rev. Methods Prim. 1 (1) (2021) 1–26.