

1 New approach and new program for analyses of false negatives-
2 contaminated data in medicine and biology

3

4 Jaroslav Flegr^{1,2*} and Petr Tureček^{1,2}

5

6 ¹ Department of Philosophy and History of Science, Faculty of Science, Charles University, Viničná 7, Prague,
7 128 43, Czech Republic

8 ² National Institute of Mental Health, Topolová 748, Klecany, 250 67, Czech Republic

9 * Corresponding author

10 Department of Philosophy and History of Science, Faculty of Science, Charles University, Viničná 7, Prague,
11 128 43, Czech Republic, E-mail: flegr@cesnet.cz; tel.: +(420) 221951821

12

13

14

15

16 Abstract

17 **Background:** No serological assay has 100% sensitivity. Statistically, the concentration of specific antibodies
18 against antigens of parasites decreases with the duration of infection. This can result in false negative outputs of
19 diagnostic tests for the subjects with old infection, e.g., for individuals infected in childhood. When a property
20 of seronegative and seropositive subjects is compared under these circumstances, the statistical tests can detect
21 no significant difference between these two groups of subjects, despite the fact that infected and noninfected
22 subjects differ. When the effect of the infection has a cumulative character and subjects with an older infection
23 (potential false negatives) are affected to a greater degree, we can even get paradoxical result of the comparison
24 – the seronegative subjects have on average lower value of certain traits, e.g. IQ, despite the infection having a
25 negative effect on the trait. A permutation test for the contaminated data, implemented, e.g., in the program
26 Trept or available as a comprehensibly commented R function in the supplement of this paper, can be used to
27 reveal and to eliminate the effect of false negatives.

28 **Methods:** We used a Monte Carlo simulation in the program R to show that the permutation test implemented in
29 the programs Trept and PTPT is a conservative test.

30 **Results:** We showed that the test could provide false negative but not false positive results if the studied
31 population contains no subpopulation of false negative subjects. We also introduced R version of the test
32 expanded by skewness analysis, which helps to estimate the proportion of false negative subjects based on the
33 assumption of equal data skewness in groups of healthy and infected individuals.

34 **Conclusions:** Based on the results of simulations and our experience with empirical studies we recommend the
35 usage of permutation test for contaminated data whenever seronegative and seropositive individuals are
36 compared.

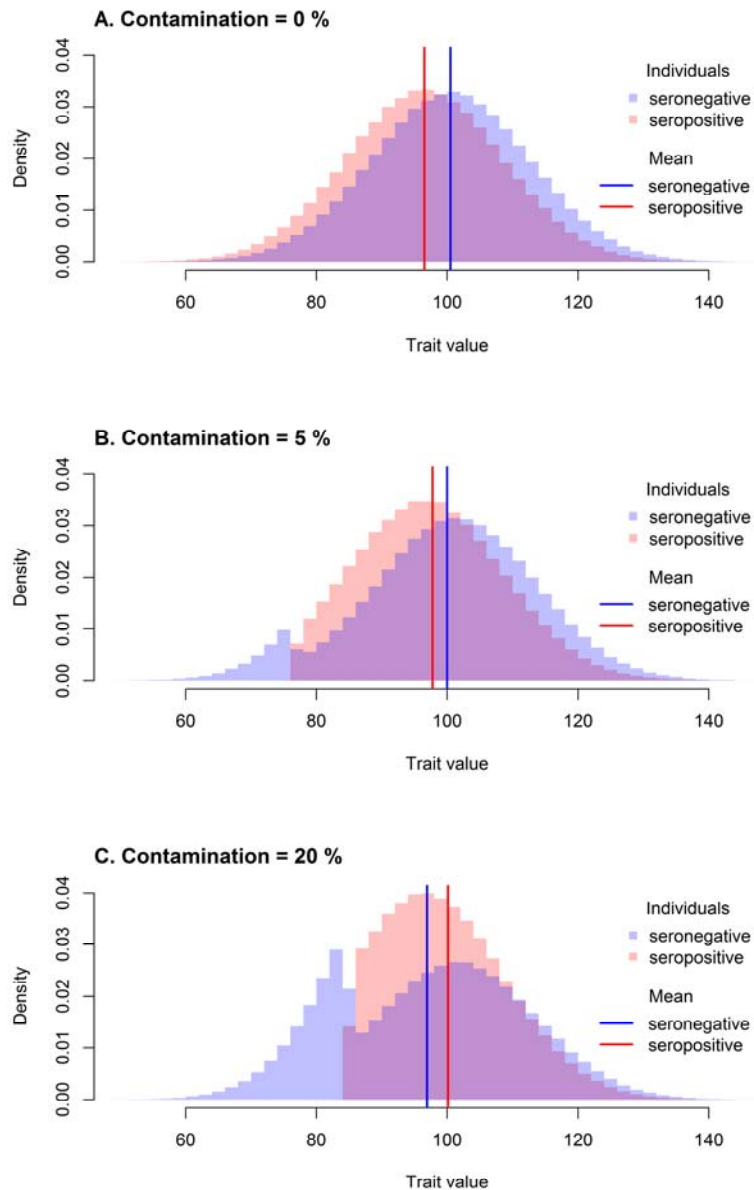
37 **Keywords:** randomisation tests; epidemiology; serology; case-control studies; specificity; sensitivity;
38 toxoplasma

39

40 Introduction

41 The reported decrease of specific antibodies with time from the onset of infection increases the risk of
42 false negative test results in subjects with old infections, e.g., in individuals infected in childhood¹⁻³. This is also
43 true for parasites that stay dormant in infected cells until the end of the life of infected hosts. Any subsample of
44 seronegative subjects could therefore be contaminated with an unknown proportion of misdiagnosed parasite-
45 positive individuals who got infected a long time ago⁴⁻⁶. This subpopulation of infected but seronegative subjects
46 could be the most influenced by the infection (Figure 1B) because of the long duration of their infection or
47 because their infection took place in early stages of their ontogenesis. This could result in a paradox (Figure 1C).
48 The seropositive subjects could have on average higher IQ scores (or higher body weight) while the intelligence
49 (or body weight) of seropositive subjects declines with the assessed length of infection (obtained from clinical
50 records or assessed by the level of antibodies).

51 **Figure 1. Exemplar distributions under 3 different contamination levels.**



52

53 *The proportion of seropositive individuals (50%), the difference between healthy and infected individuals (5) and*
54 *the standard deviation (10, corresponding to Cohen's $d = 0.5$ in non-contaminated sample) within healthy*
55 *individuals are held constant. Histogram C serves as a demonstration of the paradoxical result caused by a high*
56 *contamination when the seronegative is a lower seropositive mean trait value despite the fact that healthy*
57 *individuals score higher than infected individuals.*

58 The contamination of a parasite-free subsample with false negative individuals can be revealed and
59 eliminated by permutation tests with the reassignment of suspect cases between subsamples^{4,5}. Such permutation
60 tests can be performed using the program Treept, originally called PTPT^{7,8} modified for an analysis of data
61 contaminated with an unknown number of subjects with false negative diagnosis using the method of

62 reassignment of potentially false negative subjects⁴. This freeware program is available at
63 <http://web.natur.cuni.cz/flegr/treept.php>. The updated version of the test suited for R can be found in the
64 supplementary material of this paper in the form of comprehensibly commented R script.

65 The algorithm of the one-tailed permutation test with data reassignment is as follows: Particular
66 percentage (e.g. 5, 10, 15, 20 or 25 %) of subjects with the lowest (highest) value of the dependent variable, for
67 example IQ score, is relocated from the group of parasite-seronegative subjects to the group of the parasite-
68 seropositive subjects. Then, the difference of means of these two groups is calculated. In the next 9,999 steps, the
69 empirical values of the analysed variable are arbitrarily assigned into two groups held at the size of the original
70 seronegative and seropositive groups. The particular percentage of cases with the lowest (or highest) values of
71 the focal variable (e.g. IQ) in the pseudoseronegative group is relocated to the pseudoseropositive group, and the
72 difference between the means of the two groups is calculated. Finally, all 10,000 differences (including the one
73 calculated from non-permuted data) are sorted from highest to lowest. The percentage of the differences higher
74 or equal to that calculated on the basis of the non-permuted data is considered as the statistical significance (p) –
75 the probability of obtaining the same or higher difference between the means of two groups, if the null
76 hypothesis is correct and subjects are assigned into seropositive and seronegative groups randomly.

77 Our main aim is to show that the permutation test for contaminated data does not provide false positive
78 results, i.e., it does not return lower p than a standard permutation test if no false negative subjects exist in the
79 studied population. The second aim is to develop a new tool for the skewness analysis, which can be used to
80 estimate the approximate proportion of false-negative subjects in the studied population.

81

82 Methods and Results

83 A Monte Carlo simulation was performed with R 3.3.3. We generated a population of 150 parasite-free and 150
84 infected subjects (mean intelligence was 101.5 in the parasite-free group and 98.5 in the infected group – the
85 between-group difference was 3, the population mean intelligence was 100). Subjects were normally distributed
86 around group means with equal standard deviations (SD). We used different SDs (6, 9, 12, 15, 30) corresponding
87 to different effect sizes expressed by Cohen's d (0.5, 0.33, 0.25, 0.2, 0.1). Then we ran a standard permutation
88 test. We randomly permuted the infection status of all subjects 10,000 times and calculated a fraction of
89 permutations where the difference between two groups (pseudo-parasite-free and pseudo-parasite-infected
90 subjects) was equal to or larger than the difference between the groups in non-permuted data (p value of a
91 standard permutation test). Then, we repeated the analysis using a one-tailed permutation test for contaminated
92 data. Namely, after the generation of sets of parasite-free and parasite-infected subjects (or after the generation
93 of sets of pseudo-parasite-free and pseudo-parasite-infected subjects by permutation of the infection status), we
94 relocated 5, 10, 15, 20, 25, 30 or 50% of subjects with the lowest intelligence from the parasite-free (or pseudo-
95 parasite-free) set to the parasite-infected (or pseudo-parasite-infected) set. Again, we calculated a fraction of
96 permutations with the difference between the groups equal to or larger than the value computed for the non-
97 permuted data (p values of the permutation test for contaminated data). We used populations generated for the
98 standard permutation test (each initial population was used once for each fraction of relocated subjects). In total,
99 10,000 original populations were generated for each SD, therefore 10,000 independent permutation tests were

100 conducted for each combination of SD and each relocated fraction. The resulting p values were averaged over
 101 permutation tests with the same population SD and the same relocated fraction. The results are shown in the
 102 Table 1. With the proportion of relocated subjects, the average p-value grew for every standard deviation. The
 103 visualization of this growth can be found in Figure 2A. In this figure, the p-value of the standard permutation test
 104 was subtracted from each p-value of the permutation test for contaminated data (negative values therefore
 105 correspond to a decrease, and positive to an increase, of p-value in comparison to a standard permutation test).

106 When several exceptional data points (outliers) are present, the p-value of one or more contamination
 107 levels can be lower than p-value for 0% contamination. This is more frequent when the effect size is very small
 108 and the p-value fluctuates due to a larger impact of random noise in the data. The probability of a p-value being
 109 higher for a certain proportion of relocated data than the p-value of a standard permutation test in a particular
 110 simulation run was evaluated for each level of contamination and SD from the set of generated data described
 111 above. The results are reported in the Table 3 and shown in the Figure 3A.

112 For comparison, the same computer simulation was conducted for a population of 150 seropositive and 150
 113 seronegative individuals where 5% of seronegative individuals were false negative individuals with extremely
 114 low intelligence (example in Figure 1B). The average p-values of the permutation test for contaminated data are
 115 in Table 2. The graphical representation of the difference between a p-value for 0 % of relocated subjects and
 116 other contamination levels is represented in Figure 2B, and the probability of increase of p-value is shown in
 117 Table 4 and Figure 3B. In this case, the p-value decreases with the proportion of relocated individuals as
 118 expected.

119

120 **Table 1 Effect of relocation of hypothesized false negative subjects on the results of a permutation test if**
 121 **no such subjects exist in the population**

122

SD	Fraction of relocated subjects							
	0 %	5 %	10 %	15 %	20 %	25 %	30 %	50 %
6	.001	.001	.001	.002	.002	.002	.002	.005
9	.021	.022	.023	.024	.026	.028	.030	.044
12	.064	.066	.068	.070	.073	.076	.080	.099
15	.108	.110	.113	.116	.119	.123	.126	.148
30	.269	.270	.272	.274	.277	.280	.283	.298

123 *The table shows p-values computed with the permutation test for contaminated data when the population under*
 124 *study contains no false negative subjects. The simulation experiments were performed on populations that differ*
 125 *by variances (rows) with the relocation of different fractions of IQ-lowest individuals (columns) from the high-*
 126 *IQ (seronegative) group to the low-IQ (seropositive) group. The first column (0%) shows the (most significant)*
 127 *results of permutation tests performed without any relocation of data. For details see the Methods section. The*
 128 *fixed effect was 3 IQ points. The population size was 300, and the proportion of seropositive individuals in the*
 129 *original sample (0% relocation) was 0.5.*

130

131 **Table 2** Effect of the relocation of hypothesized false negative subjects on the results of a permutation test if 5%
 132 of such individuals is present in seronegative group

SD	Fraction of relocated subjects							
	0 %	5 %	10 %	15 %	20 %	25 %	30 %	50 %
6	.058	.035	.022	.017	.015	.014	.014	.017
9	.273	.208	.157	.131	.116	.107	.102	.100
12	.446	.371	.303	.263	.236	.219	.209	.191
15	.565	.491	.418	.371	.337	.315	.302	.274
30	.775	.721	.659	.610	.571	.542	.523	.467

133 *The table shows p-values computed with the permutation test for contaminated data when the seronegative*
 134 *group contains 5% of false negative subjects. The simulation experiments were performed on populations that*
 135 *differ by variances (rows) with relocation of different fractions of IQ-lowest individuals (columns) from the high-*
 136 *IQ (seronegative) group to the low-IQ (seropositive) group. The first column (0%) shows the (least significant)*
 137 *results of permutation tests performed without any relocation of data. For details see the Methods section. The*
 138 *fixed effect was 3 IQ points. The population size was 300, and the proportion of seropositive individuals in the*
 139 *original sample (0% relocation) was 0.5.*

140

141 **Table 3.** The probability of a p-value being higher in a particular simulation run than a p-value with 0 % of
 142 relocated individuals. No false negative subjects are present in the population.

SD	Fraction of relocated subjects						
	5 %	10 %	15 %	20 %	25 %	30 %	50 %
6	.18	.23	.27	.31	.35	.39	.58
9	.45	.50	.54	.58	.61	.64	.75
12	.50	.54	.57	.60	.62	.64	.71
15	.51	.55	.57	.59	.61	.63	.68
30	.50	.52	.53	.54	.55	.56	.59

143 *The probability that p-value will increase for specified fraction of relocated individuals in a particular*
 144 *simulation run as compared to 0 % of relocated seronegative individuals. The simulated population are identical*
 145 *to the population represented in Table 1. The graphical summary can be found in Figure 3A. The relatively*
 146 *small numbers in first row are caused by the fact that in many simulation runs p-values remained <.0001 for a*
 147 *small fraction of relocated subjects when the effect size is relatively large.*

148

149

150

151

152 **Table 4.** The probability of p-value being higher in particular simulation run than p-value with 0 % of relocated
153 individuals. 5% of seronegative group are false negative individuals.

SD	Fraction of relocated subjects						
	5 %	10 %	15 %	20 %	25 %	30 %	50 %
6	.00	.00	.00	.00	.00	.00	.06
9	.00	.00	.00	.00	.00	.00	.03
12	.00	.00	.00	.00	.00	.00	.02
15	.00	.00	.00	.00	.00	.00	.02
30	.00	.00	.00	.00	.00	.00	.01

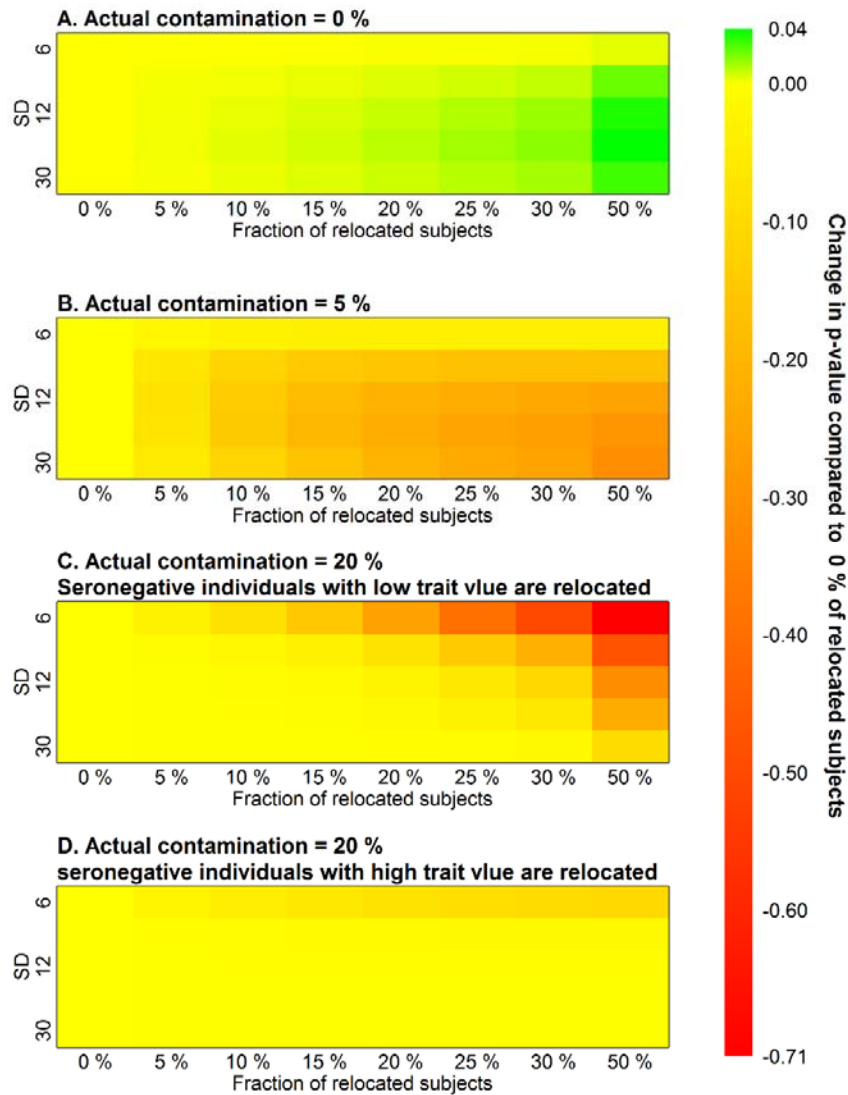
154 *The probability that a p-value will increase for a specified fraction of relocated individuals in a particular*
155 *simulation run as compared to 0 % of relocated seronegative individuals. The simulated population are identical*
156 *to those represented in Table 2. The graphical summary can be found in Figure 3B.*

157

158 Two equivalent simulations were run to demonstrate the permutation test for contaminated data on the
159 paradoxical dataset with a high proportion of false negative individuals. The first population of 150 seropositive
160 and 150 seronegative individuals where 20% of seronegative subjects were false negative individuals with
161 extremely low intelligence (Figure 1C). A similar one-tailed permutation test as in previous simulations was run
162 as it was hypothesised that the average trait value of the healthy group is actually higher despite the paradoxical
163 situation. The graphical representations of the results are in Figure 2C and Figure 3C. The second test with the
164 same sample generation algorithm (150 seropositive, 150 seronegative, 20% false negative) was set to follow the
165 default setting of the permutation test for contaminated data, which assumes the non-paradoxical situation and
166 therefore relocates seronegative individuals with high trait value, thus widening the gap between the groups.
167 Yielded p-value of one-tailed permutation test is then the proportion of random samples after relocation where
168 the difference between groups (seronegative - seropositive) was lower than in the original sample (Figure 2D and
169 Figure 3D). In both simulation tests on a sample with 20% contamination, the p-value of respective one-tailed
170 permutation test decreased, so this sample was clearly distinguishable from the case in which no false negative
171 subjects were present.

172 The appropriate direction of subject relocation can be determined on the basis of a skewness analysis of
173 the original sample, which is available in the R version of the test⁹ if a parameter skewness.analysis is set to
174 TRUE. The skewness analysis and its usage for the assessment of group mean order as well as the contamination
175 level estimation is described in the Appendix of this paper. Using a two-tailed test is also worth consideration in
176 this case. The p-value is then declining with the proportion of relocated individuals in all cases where false
177 negative individuals are present (in well identified paradoxical situations only after the group means change their
178 order into the right direction).

179 **Figure 2: Heatmap of the average difference between the p-value of standard permutation test and p-**
180 **value of the respective permutation test for contaminated data.**



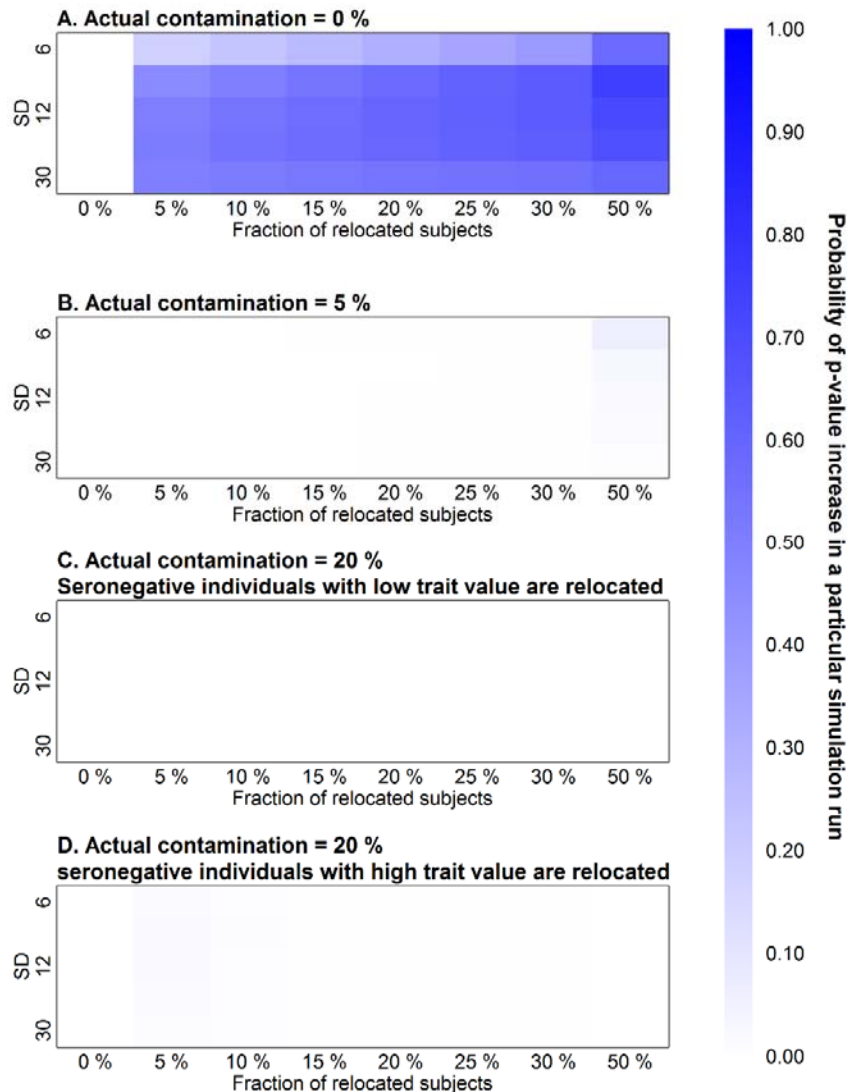
181

182 *One-tailed tests were used. The p-value increases with the fraction of relocated individuals if no actual false*
183 *negative individuals are present (A) and decreases if the sample is contaminated (B, C). This is true even if the*
184 *wrong relocation direction is employed due to a paradoxical switch in the order of group means (D).*

185

186

187 **Figure 3: Heatmap of probability of p-value increase in particular simulation run.**



188

189 *The p-value does not increase in 100% of non-contaminated samples when a permutation test for contaminated*
190 *data is used, but the probability that it happens is very high compared to samples in which false negative*
191 *individuals occur.*

192

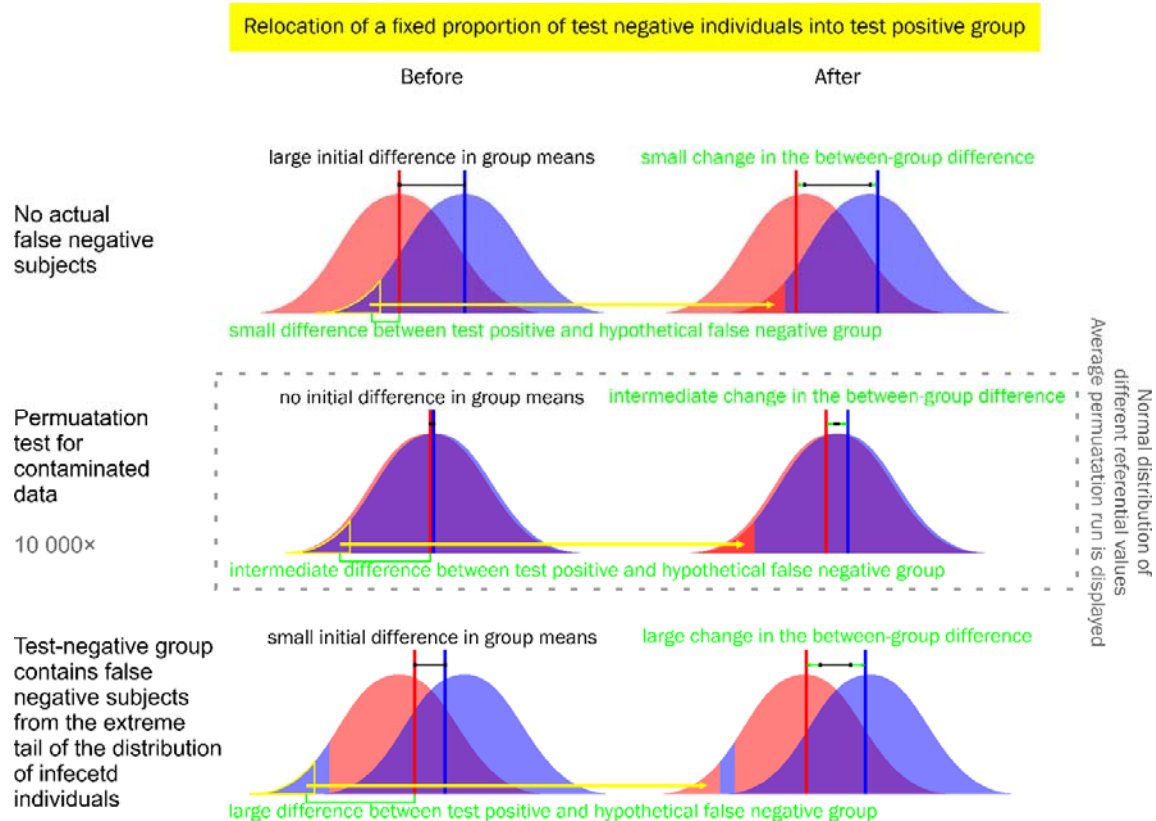
193 Conclusions

194 The results of simulation showed that the permutation test for contaminated data does not provide more
195 significant results than a standard permutation test if the experimental data does not contain a subpopulation of
196 false negative subjects. This test is conservative when its usage is not necessary and allows one to avoid false
197 negative results in the case of data contamination. This is due to the higher difference between the relocated
198 seronegative and the original seropositive group in the presence of false negative data. The referential set of
199 permutations with relocation remains the same in both cases, while the relative change in inter-group difference

200 after relocation maintains an intermediate position between those two options (see Figure 4). Therefore, the
 201 positive result of this test, i.e. lowering the p-value with the growth of the proportion of relocated individuals,
 202 itself supports the hypothesis that the set of seemingly parasite-free subjects contains false negative subjects,
 203 who, most probably, have become infected a long time ago or in very young age.

204

205 Figure 4. Graphical demonstration of the intermediate position of referential permutations with relocation
 206 between empirical cases of relocation of seronegative healthy subjects and false negative infected subjects.



207

208

209 *The increase in p-value in the case of non-contaminated data is much smaller than the increase caused by*
 210 *possible contamination, which can completely wipe out the actual inter-group difference or even cause a*
 211 *paradoxical switch of the group mean order. (See Table 5 or the position of 0 in the legend of Figure 2.)*

212

213

214 **Table 5.** Risk associated with different combinations of data and used permutation tests

		Test	
		regular permutation test	permutation test for contaminated data
Data	non-contaminated	No risk	Small risk of false negative results
	contaminated	High risk of false negative or paradoxical results	No risk

215 *The pressure to avoid a high risk associated with regular permutation test will lead us to the universal utilization*
 216 *of permutation test for contaminated data whenever properties of seropositive and seronegative subjects are*
 217 *compared. When we conduct a skewness analysis for contaminated data (see the Appendix) prior to the*
 218 *permutation test, we can lower the risks further by justification of regular permutation test or informed setting of*
 219 *relocated fractions of seronegative individuals in permutation test for contaminated data.*

220 Based on the theoretical grounding described above and our experience with the research of two
 221 unrelated species of parasites, *Toxoplasma gondii*^{5, 10} and human cytomegalovirus⁶, we strongly recommend the
 222 usage of permutation tests for contaminated data⁹ whenever any properties of parasite-infected and parasite-free
 223 individuals are compared.

224

225 **Conflict of interest statement**

226 The authors declare to have no conflicts of interest.

227 **Acknowledgements**

228 We would like to thank Charlie Lotterman for his help with the final version of the paper.

229

230 **Funding**

231 This work has been supported by Czech Science Foundation 18-13692S and Charles University Research Centre
 232 program No. 204056.

233

234 **Appendix 1**

235 **Estimation of the fraction of false negatives by skewness analysis**

236 The estimation of the actual contamination level is very difficult to discern and should be investigated more in
237 future research. For now, we can seek assistance in a skewness analysis, which compares the skewness of trait
238 value distribution in seropositive and seronegative groups. Skewness is defined as third standardized moment
239 measuring the asymmetry of the probability distribution. We assume that healthy and infected individuals have
240 an equally skewed trait distribution. This assumption is violated if false negative individuals are recruited from
241 one of extreme tails of the distribution of infected individuals. If, for example, infected individuals with the
242 lowest trait value are identified as seronegative (as seen in Figure 1B), the skewness of seropositive individuals
243 becomes more positive and the skewness of seronegative individuals more negative. The exact opposite is true if
244 individuals from the upper tail of the distribution are misdiagnosed as negative. The skewness comparison
245 (available as a function in supplementary R script) of contaminated data compares Fisher-Pearson coefficient of
246 skewness of seropositive and seronegative groups under different hypothesised contamination levels and returns
247 the skewness values for each fraction of relocated subjects, p-values of the difference between them based on
248 permutation test, the interval where the group skewness is not significantly different and a proportion of
249 relocated seronegative individuals at which the difference between group skewness was smallest (i.e. the one the
250 generated the most similarly skewed groups). This value generally underestimates the actual contamination, but
251 any amendments would require additional assumptions about the distribution of healthy/infected individuals,
252 which would not be necessarily met in empirical data. **Now we recommend the conduction a skewness
253 comparison prior to the evaluation of the between-group difference and then the conduction of a
254 permutation test for contaminated data for contamination levels between 0 and upper border of the
255 interval, where the difference between group skewness was not significant.** We observed that the actual level
256 of contamination in simulated data, where we can control the contamination level, rather closely matches the
257 upper level of the similar-skewness interval due to the fact that the distributions of healthy and infected subjects
258 largely overlap, and the extreme tail of seronegative distribution contains also extreme healthy individuals which
259 are relocated prior to actual false negative individuals. For the same reason, however, we can suggest that the
260 between-group difference for the relocated fraction where the group skewness are most similar (described above)
261 closely matches the actual between-group difference in non-contaminated populations without false negative
262 subjects.

263 The difference between skewness coefficients in seropositive and seronegative groups in the original
264 sample without relocated individuals can also be evaluated in the R version of the permutation test for
265 contaminated data⁹ (set skewness.analysis to TRUE). This analysis allows one to appropriately assess whether
266 the seronegative group includes false negative subjects from the extreme tail of the distribution of infected
267 individuals. By default, the permutation test for contaminated data assumes that the observed order of mean
268 values of seropositive and seronegative groups accurately reflects the state of things in correctly determined
269 healthy and infected groups. Therefore, the function will gradually relocate individuals from the lower tail of the
270 distribution if the seronegative mean trait value is higher than the seropositive mean and vice versa (this can be
271 changed by the parameter higher.healthy). If we do not alter default setting in paradoxical situations (Figure 1C),
272 in which the order of group means was changed due to contamination, the test algorithm will increase the

273 difference between the groups by relocating healthy individuals from the upper tail of distribution of
274 seronegative subjects. The p-value will most likely decrease with the growing fraction of relocated individuals,
275 as in other cases where false negative individuals are present. This might lead to a radical misinterpretation of
276 the data (confirmation of the assumption of higher trait value in group of infected individuals) if attention is not
277 paid to the skewness analysis. The skewness analysis of the original sample is not fooled as easily since
278 mismatching the extreme tail of infected individuals as seronegative will alter the skewness of both groups
279 substantially. Under an extreme proportion of false negatives, the skewness of both groups might actually be
280 shifted in the same direction (e.g. positive in cases similar to the example in Figure 1C). However, the skewness
281 of seropositive group will be still substantially more deflected than the skewness of the seronegative group, so
282 the skewness analysis will return reliable results.

283

284 **References**

- 285 1. Celik AD, Yulugkural Z, Kilincer C, Hamamcioglu MK, Kuloglu F, Akata F. Negative serology:
286 could exclude the diagnosis of brucellosis? *Rheumatol Int* 2012; **32**: 2547-9.
- 287 2. El-Sherif A, Elbahrawy A, Aboelfotoh A, et al. High false-negative rate of anti-HCV among
288 Egyptian patients on regular hemodialysis. *Hemodial Int* 2012; **16**: 420-7.
- 289 3. Brown SL, Hansen SL, Langone JJ. Role of serology in the diagnosis of Lyme disease. *Jama-J*
290 *Am Med Assoc* 1999; **282**: 62-6.
- 291 4. Flegr J, Havlíček J. Changes in the personality profile of young women with latent
292 toxoplasmosis. *Folia Parasitol* 1999; **46**: 22-8.
- 293 5. Flegr J, Hrdá Š, Kodym P. Influence of latent 'asymptomatic' toxoplasmosis on body weight of
294 pregnant women. *Folia Parasitol* 2005; **52**: 199-204.
- 295 6. Chvatalova V, Sebankova B, Hrbackova H, Turecek P, Flegr J. Differences in cognitive
296 functions between cytomegalovirus-infected and cytomegalovirus-free university students: a case
297 control study. *Sci Rep* 2018; **8**.
- 298 7. Flegr J, Záboj P. PTPT, the freeware program for permutation testing concordance between
299 phylogeny and the distribution of phenetic traits. *Acta Soc Zool Bohem* 1997; **61**: 91-5.
- 300 8. Flegr J, Záboj P, Vaňáčková Š. Correlation between aerobic and anaerobic resistance to
301 metronidazole in trichomonads: application of a new computer program for permutation tests.
302 *Parasitol Res* 1998; **84**: 590-2.
- 303 9. Flegr J, Turecek P. Permutation test for contaminated data and skewness analysis. *Figshare*
304 2019.
- 305 10. Flegr J, Kodym P, Tolarová V. Correlation of duration of latent *Toxoplasma gondii* infection
306 with personality changes in women. *Biol Psychol* 2000; **53**: 57-68.

307

308

309

310 **Supplements (R scripts)**

311 (Jaroslav Flegr and Petr Tureček, New approach and new program for analyses of false negatives-contaminated
312 data in medicine and biology)

313

314 **Supplement 1**

315 **Permutation test for contaminated data**

316 #####

317 ####hit ctrl+a and ctrl+r to install the function####

318 #####

319

320 #This script contains contamination_perm_test() function

321 #contamination_perm_test(trait,identification,percentages=c(0,5,10),higher.

322 healthy=(mean(trait[identification==F])>mean(trait[identification==T])),run

323 s=10000,skewness.analysis=F)

324

325 #Arguments are described below.

326 #It is necessary to define funtion calculating skewness index first:

327

328 ##Fuction that returns Fisher-Pearson coefficient of skewness.

329 ##Input is a vector of numerical values.

330

331 FPskewness<-function(x){

332 return((sum((x-mean(x))^3)/length(x))/((sqrt(sum((x-

333 mean(x))^2)/length(x))^3))

334 }

335

336

337 ###contamination_perm_test

338 ###Function that delegates the parameters to either one-tailed or two-

339 tailed tests described below

340

```
341 contamination_perm_test<-
342 function(trait,identification,percentages=c(0,5,10),higher.healthy=(mean(trait[identification==F])>mean(trait[identification==T])),runs=10000,two.tailed=F,skewness.analysis=F){
343
344
345   if(two.tailed==F){
346     contamination_perm_test_one(trait=trait,identification=identification,percentages=percentages,higher.healthy=higher.healthy,runs=runs,skewness.analysis=skewness.analysis)
347
348   }else{
349     contamination_perm_test_two(trait=trait,identification=identification,percentages=percentages,higher.healthy=higher.healthy,runs=runs,skewness.analysis=skewness.analysis)
350
351   }
352
353 }
354
355
356 ###This Function works with following arguments:
357
358 ###trait - Numerical vector of trait values
359
360 ###identification - Logical vector of assumed presence (T) or absence (F) of infection
361
362 ###percentages - Numerical vector of percentages of false negative amongst negative subjects (contamination levels) for which the permutation test for contaminated data will be run.
363
364 ###two.tailed - Specifies the version of the test, two.tailed=F is the default.
365
366 ###higher healthy - Logical. Indicates whether we assume the healthy individuals to show higher (T) or lower (F) trait values. When not specified, the script assumes this relationship based of group means with no hypothesised contamination.
367
368 #####It allows us to use the difference between the groups (not in absolute values) in permutation test. In this scenario the seropositive group mean is subtracted from seronegative group mean and the one-tailed permutation test is conducted accordingly.
369
370
371
372
373 ###runs - Number of resamplings used in permutation test
374
375
376 ###One-tailed version of the test
377
```

```
378 contamination_perm_test_one<-
379 function(trait,identification,percentages=c(0,5,10),higher.healthy=(mean(trait[identification==F])>mean(trait[identification==T])),runs=10000,skewness
380 .analysis=F){
381
382
383 if(length(trait)!=length(identification)){
384 stop("The vectors of trait values and infection indication are of different
385 lengths.")
386 }
387
388 higher<-(mean(trait[identification==F])>mean(trait[identification==T]))
389 set.higher<-higher.healthy
390
391 orig.means<-tapply(trait,identification,mean)
392 orig.means<-data.frame(orig.means)
393
394 names(orig.means)<-"Original mean values"
395 rownames(orig.means)[which(rownames(orig.means)=="FALSE")]<-"Identified as
396 healthy"
397 rownames(orig.means)[which(rownames(orig.means)=="TRUE")]<-"Identified as
398 infected"
399
400 higher.report<-ifelse(higher==T,
401 "In original sample, individuals identified as healthy showed higher
402 \naverage trait value.",
403 "In original sample, individuals identified as infected showed higher
404 \naverage trait value."
405 )
406
407 concord<-ifelse(higher==set.higher,"Consequently,","Despite that,")
408
409 set.report1<-paste(concord,ifelse(set.higher==T,
410 "healthy individuals were hypothesised to have higher \naverage trait value
411 in a contamination-free sample. \n",
```



```
412 "infected individuals were hypothesised to have higher \naverage trait
413 value in a contamination-free sample. \n"
414 ))
415
416 set.report2<-paste(iffelse(set.higher==T,
417 "\nFor each contamination level respective proportion of seronegative
418 \nindividuals with lowest trait value was relabeled as seropositive \nin
419 original sample as well as in each permutation test run.",
420 "\nFor each contamination level respective proportion of seronegative
421 \nindividuals with highest trait value was relabeled as seropositive \nin
422 original sample as well as in each permutation test run."
423 ))
424
425 trait<-c(trait[identification==F],trait[identification==T])
426 infected<-sort(identification)
427
428 count.healthy<-sum(!identification)
429 count.infected<-sum(identification)
430
431 Nperc<-length(percentages)
432
433 vector.ident<-list()
434
435 for(i in 1:Nperc){
436   reassign<-round(count.healthy*(percentages[i]/100))
437   identification<-c(rep(F,count.healthy-
438   reassign),rep(T,count.infected+reassign))
439   vector.ident[[i]]<-identification
440 }
441
442 which.test<-paste("One-tailed permutation test for contaminated data was
443 executed. \nProportion of differences (mean of non-infected - mean of
444 infected)",
445   iffelse(set.higher==T,"\nhigher","\nlower"),
```

```
446 "than the observed difference is returned as an equivalent \nof p-
447 value.\n",collapse=" ")
448
449 #Sorts healthy individuals to indicate possible false-negatives
450 trait2<-
451 c(sort(trait[infected==F],decreasing=higher.healthy),trait[infected==T])
452
453 dist.reals<-1:Nperc
454 names(dist.reals)<-paste(as.character(percentages), "%")
455
456 contamination<-paste(as.character(percentages), "%")
457 names(contamination)<-paste(as.character(percentages), "%")
458
459 mean.healthy<-dist.reals
460 mean.infected<-dist.reals
461
462 sd.healthy<-dist.reals
463 sd.infected<-dist.reals
464
465 N.healthy<-dist.reals
466 N.infected<-dist.reals
467
468 mean.dist.perm<-dist.reals
469 p.vals.perm<-dist.reals
470
471 #Compute group means in non-permuted sample
472 for(i in 1:Nperc){
473 mean.healthy[i]<-mean(trait2[vector.ident[[i]]==F])
474 mean.infected[i]<-mean(trait2[vector.ident[[i]]==T])
475 sd.healthy[i]<-sd(trait2[vector.ident[[i]]==F])
476 sd.infected[i]<-sd(trait2[vector.ident[[i]]==T])
```

```
477 N.healthy[i]<-sum(vector.ident[[i]]==F)
478 N.infected[i]<-sum(vector.ident[[i]]==T)
479 dist.reals[i]<-(mean(trait2[vector.ident[[i]]==F])-
480 mean(trait2[vector.ident[[i]]==T]))
481 }
482
483 cohen<-
484 abs(dist.reals)/((sd.healthy*N.healthy+sd.infected*N.infected)/(N.healthy+N
485 .infected))
486
487 #Skewness computation
488 skewness.healthy<-FPskewness(trait[infected==F])
489 skewness.infected<-FPskewness(trait[infected==T])
490
491 skew.diff<-abs(skewness.healthy-skewness.infected)
492
493 skewness<-c(skewness.healthy,skewness.infected)
494 skewness<-data.frame(skewness)
495
496 names(skewness)<-"Fisher-Pearson coefficient of skewness"
497 rownames(skewness)<-c("Identified as healthy","Identified as infected")
498
499 signs<-sign(dist.reals)
500
501 #Permutation test with skewness add-on
502 perm.dist<-array(NA,dim=c(Nperc,runs))
503 rand.skew<-NA
504
505 for(run in 1:runs){
506 trait2<-sample(trait)
507 rand.skew[run]<-abs(FPskewness(trait2[infected==F])-
508 FPskewness(trait2[infected==T]))
```

```
509
510 trait2<-
511 c(sort(trait2[infected==F],decreasing=higher.healthy),trait2[infected==T])
512
513 for(i in 1:Nperc){
514 perm.dist[i,run]<-mean(trait2[vector.ident[[i]]==F]) -
515 mean(trait2[vector.ident[[i]]==T])
516 }
517 }
518
519 skew.p<-sum(rand.skew>skew.diff)/runs
520
521 skew.higher<-ifelse(skewness.healthy>skewness.infected,"test-
522 negative","test-positive")
523 skew.guess.higher<-ifelse(skewness.healthy>skewness.infected,FALSE,TRUE)
524 healthy.positive<-ifelse(skewness.healthy>0,TRUE,FALSE)
525 infected.positive<-ifelse(skewness.infected>0,TRUE,FALSE)
526 skew.sig<-ifelse(skew.p<0.05,TRUE,FALSE)
527
528 skew.message<-paste(
529 ifelse(healthy.positive==infected.positive,
530 paste(
531 "The distribution of trait value was",
532 ifelse(healthy.positive,"positively","negatively"),
533 "skewed \nin both groups.",
534 "The Fisher-Pearson coefficient of skewness \nwas higher
535 in",skew.higher,"group.")
536 ,
537 paste("The distribution of individuals identified as healthy \nwas skewed",
538 ifelse(healthy.positive,"positively","negatively"),
539 "the distribution of individuals \nidentified as infected",
540 ifelse(infected.positive,"positively","negatively."))
```

```
541 )
542 ,
543 paste("\n\nThe difference between the coefficients of skewness was",
544 ifelse(skew.sig," \nstatistically significant.\n",", \nhowever, not
545 statistically significant.\n"),
546 "(Two-tailed permutation test of skewness difference \non ",runs," runs was
547 executed.)",sep="")
548 ,
549 ifelse(skew.sig==FALSE,
550 "\n\nThis might question the assumption of data contamination \nsince we
551 would expect a difference in skewness between \nthe groups in contaminated
552 data. \nProceed with caution.",
553 paste("\n\nThis supports the assumption of data contamination.",
554 "\nBased on the difference in skewness we would assume \ncontamination of
555 healthy group by false negative \nsubjects from the",
556 ifelse(skew.higher=="test-positive","lower","upper"),
557 "tail of the distribution \nof infected individuals, which would lead to
558 overall",
559 ifelse(skew.higher=="test-positive","\ndecrease","\nincrease"),
560 "of test-negative group mean.))
561 ,
562 ifelse(skew.sig==FALSE,"",
563 paste(ifelse(set.higher==skew.guess.higher,
564 paste("\n\nThe skewness analysis brings further support to the hypothesis
565 \nof",
566 ifelse(set.higher,"higher","lower"),
567 "mean in non-contaminated group of healthy \nindividuals, which was used in
568 current permutation test \nfor contaminated data.\n"),
569 paste("\n\nThe skewness analysis, however, does not support the hypothesis
570 \nof",
571 ifelse(set.higher,"higher","lower"),
572 "mean in non-contaminated group of healthy \nindividuals, which was used in
573 current permutation test \nfor contaminated data. Proceed with caution.\n")
574 )))
575 )
```

```
576
577 skewness<-rbind(skewness [1] , "", skew.p)
578
579 rownames (skewness) [c (3,4) ]<-c ("", "p-value")
580
581 skewness [c (1,2,4) ,1]<-format (round (as.numeric (skewness [c (1,2,4) ,1] ) ,3))
582
583 mean.dist.perm<-rowMeans (perm.dist)
584
585 if (higher.healthy==T) {
586   for(i in 1:Nperc) {
587     p.vals.perm[i] <-sum(dist.reals[i] <perm.dist [i,] )/runs
588   }
589 }else{
590
591   for(i in 1:Nperc) {
592     p.vals.perm[i] <-sum(dist.reals[i] >perm.dist [i,] )/runs
593   }
594 }
595
596 mean.healthy<-format (round (mean.healthy,2))
597 mean.infected<-format (round (mean.infected,2))
598 sd.healthy<-format (round (sd.healthy,2))
599 sd.infected<-format (round (sd.infected,2))
600 cohen<-format (round (cohen,2))
601 dist.reals<-format (round (dist.reals,2))
602
603 mean.dist.perm<-format (round (mean.dist.perm,2))
604 p.vals.perm<-format (round (p.vals.perm,3))
605
```

```
606
607 res.table<-
608 rbind(contamination,mean.healthy,mean.infected,dist.reals,mean.dist.perm,sd
609 .healthy,sd.infected,N.healthy,N.infected,cohen,p.vals.perm)
610 res.table<-as.data.frame(res.table)
611
612 colnames(res.table)<-NULL
613 rownames(res.table)<-c("contamination","non-infeceted mean","infected
614 mean","mean difference","expected difference","non-infeceted SD","infected
615 SD","non-infeceted N","infected N","Cohen's d","p-value")
616
617 final.message<-ifelse(all(signs>0),
618 "\nThe mean difference was positive in all \nhypothesised contamination
619 levels.",
620 ifelse(all(signs<0),
621 "\nThe mean difference was negative in all \nhypothesised contamination
622 levels.",
623 ifelse(signs[1]>0,
624 "\nThe mean difference started as positive, but turned negative \nwith
625 growing hypothesised contamination level. \nThe results should be
626 interpreted with caution. \nRunning the test that assumess the opposite
627 relationship \nbetween group means (higher.healthy=T) or a two tailed test
628 \nis worth consideration.",
629 "\nThe mean difference started as negative, but turned positive \nwith
630 growing hypothesised contamination level. \nThe results should be
631 interpreted with caution. \nRunning the test that assumess the opposite
632 relationship \nbetween group means (higher.healthy=F) or a two tailed test
633 \nis worth consideration."
634 )))
635
636 cat("\nSample characteristics:\n")
637 print(orig.means)
638 cat(paste("\n",higher.report,"\n\n",sep=""))
639 cat(set.report1)
640
641 if(skewness.analysis==T) {
```

```
642   cat ("\n\nSkewness report:\n")
643   print(skewness)
644   cat ("\n")
645   cat(skew.message)
646 }
647
648   cat ("\n\nPermutation test for contaminated data:\n")
649   cat(paste(set.report2, "\n\n", sep=""))
650   cat(which.test)
651   cat(paste("\n", runs, "sample permutations were performed.\n"))
652   print(res.table)
653   cat(paste(final.message, "\n\n", sep=""))
654
655   results<-
656   list(orig.means, higher.report, set.report1, skewness, skew.message, set.report2
657   , which.test, res.table, final.message)
658
659   return(invisible(results))
660 }
661
662
663
664   ###Two-tailed version of the test:
665
666   contamination_perm_test_two<-
667   function(trait, identification, percentages=c(0,5,10), higher.healthy=(mean(tr
668   ait[identification==F])>mean(trait[identification==T])), runs=10000, skewness
669   .analysis=F){
670
671   if(length(trait)!=length(identification)){
672   stop("The vectors of trait values and infection indication are of different
673   lengths.")
674   }
```



```
675
676 higher<- (mean(trait [identification==F]) >mean(trait [identification==T]))
677 set.higher<-higher.healthy
678
679 orig.means<-tapply(trait,identification,mean)
680 orig.means<-data.frame(orig.means)
681
682 names(orig.means)<-"Original mean values"
683 rownames(orig.means)[which(rownames(orig.means)=="FALSE")]<-"Identified as
684 healthy"
685 rownames(orig.means)[which(rownames(orig.means)=="TRUE")]<-"Identified as
686 infected"
687
688 higher.report<-ifelse(higher==T,
689 "In original sample, individuals identified as healthy showed higher
690 \naverage trait value.",
691 "In original sample, individuals identified as infected showed higher
692 \naverage trait value."
693 )
694
695 concord<-ifelse(higher==set.higher,"Consequently,","Despite that,")
696
697 set.report1<-paste(concord,ifelse(set.higher==T,
698 "healthy individuals were hypothesised to have higher \naverage trait value
699 in a contamination-free sample. \n",
700 "infected individuals were hypothesised to have higher \naverage trait
701 value in a contamination-free sample. \n"
702 ))
703
704 set.report2<-paste(ifelse(set.higher==T,
705 "\nFor each contamination level respective proportion of seronegative
706 \nindividuals with lowest trait value was relabeled as seropositive \nin
707 original sample as well as in each permutation test run.",
```

```
708 "\nFor each contamination level respective proportion of seronegative
709 \nindividuals with highest trait value was relabeled as seropositive \nin
710 original sample as well as in each permutation test run."
711 ))
712
713 trait<-c(trait[identification==F],trait[identification==T])
714 infected<-sort(identification)
715
716 count.healthy<-sum(!identification)
717 count.infected<-sum(identification)
718
719 Nperc<-length(percentages)
720
721 vector.ident<-list()
722
723 for(i in 1:Nperc){
724   reassign<-round(count.healthy*(percentages[i]/100))
725   identification<-c(rep(F,count.healthy-
726     reassign),rep(T,count.infected+reassign))
727   vector.ident[[i]]<-identification
728 }
729
730 which.test<-paste("Two-tailed permutation test for contaminated data was
731 executed. \nProportion of differences (in absolute value)",
732 "higher than the \nobserved difference is returned as an equivalent of p-
733 value.\n",collapse=" ")
734
735 #Sorts healthy individuals to indicate possible false-negatives
736 trait2<-
737 c(sort(trait[infected==F],decreasing=higher.healthy),trait[infected==T])
738
739 dist.reals<-1:Nperc
740 names(dist.reals)<-paste(as.character(percentages), "%")
```

```
741
742 contamination<-paste(as.character(percentages), "%")
743 names(contamination)<-paste(as.character(percentages), "%")
744
745 mean.healthy<-dist.reals
746 mean.infected<-dist.reals
747
748 sd.healthy<-dist.reals
749 sd.infected<-dist.reals
750
751 N.healthy<-dist.reals
752 N.infected<-dist.reals
753
754 mean.dist.perm<-dist.reals
755 p.vals.perm<-dist.reals
756
757 #Compute group means in non-permuted sample
758 for(i in 1:Nperc){
759 mean.healthy[i]<-mean(trait2[vector.ident[[i]]==F])
760 mean.infected[i]<-mean(trait2[vector.ident[[i]]==T])
761 sd.healthy[i]<-sd(trait2[vector.ident[[i]]==F])
762 sd.infected[i]<-sd(trait2[vector.ident[[i]]==T])
763 N.healthy[i]<-sum(vector.ident[[i]]==F)
764 N.infected[i]<-sum(vector.ident[[i]]==T)
765 dist.reals[i]<-abs((mean(trait2[vector.ident[[i]]==F]) -
766 mean(trait2[vector.ident[[i]]==T])))
767 }
768
769 cohen<-
770 abs(dist.reals)/((sd.healthy*N.healthy+sd.infected*N.infected)/(N.healthy+N
771 .infected))
772
```

```
773 #Skewness computation
774 skewness.healthy<-FPskewness(trait[infected==F])
775 skewness.infected<-FPskewness(trait[infected==T])
776
777 skew.diff<-abs(skewness.healthy-skewness.infected)
778
779 skewness<-c(skewness.healthy,skewness.infected)
780 skewness<-data.frame(skewness)
781
782 names(skewness)<-"Fisher-Pearson coefficient of skewness"
783 rownames(skewness)<-c("Identified as healthy","Identified as infected")
784
785 signs<-sign(dist.reals)
786
787 #Permutation test with skewness add-on
788 perm.dist<-array(NA,dim=c(Nperc,runs))
789 rand.skew<-NA
790
791 for(run in 1:runs){
792   trait2<-sample(trait)
793   rand.skew[run]<-abs(FPskewness(trait2[infected==F]) -
794     FPskewness(trait2[infected==T]))
795
796   trait2<-
797     c(sort(trait2[infected==F],decreasing=higher.healthy),trait2[infected==T])
798
799   for(i in 1:Nperc){
800     perm.dist[i,run]<-abs(mean(trait2[vector.ident[[i]]==F]) -
801       mean(trait2[vector.ident[[i]]==T]))
802   }
803 }
804
```

```
805 skew.p<-sum(rand.skew>skew.diff)/runs
806
807 skew.higher<-ifelse(skewness.healthy>skewness.infected,"test-
808 negative","test-positive")
809 skew.guess.higher<-ifelse(skewness.healthy>skewness.infected,FALSE,TRUE)
810 healthy.positive<-ifelse(skewness.healthy>0,TRUE,FALSE)
811 infected.positive<-ifelse(skewness.infected>0,TRUE,FALSE)
812 skew.sig<-ifelse(skew.p<0.05,TRUE,FALSE)
813
814 skew.message<-paste(
815 ifelse(healthy.positive==infected.positive,
816 paste(
817 "The distribution of trait value was",
818 ifelse(healthy.positive,"positively","negatively"),
819 "skewed \nin both groups.",
820 "The Fisher-Pearson coefficient of skewness \nwas higher
821 in",skew.higher,"group.")
822 ,
823 paste("The distribution of individuals identified as healthy \nwas skewed",
824 ifelse(healthy.positive,"positively","negatively"),
825 "the distribution of individuals \nidentified as infected",
826 ifelse(infected.positive,"positively","negatively."))
827 )
828 ,
829 paste("\n\nThe difference between the coefficients of skewness was",
830 ifelse(skew.sig," \nstatistically significant.\n"," \nhowever, not
831 statistically significant.\n"),
832 "(Two-tailed permutation test of skewness difference \non ",runs," runs was
833 executed.)",sep="")
834 ,
835 ifelse(skew.sig==FALSE,
```

```
836 "\n\nThis might question the assumption of data contamination \nsince we
837 would expect a difference in skewness between \nthe groups in contaminated
838 data. \nProceed with caution.",
839 paste("\n\nThis supports the assumption of data contamination.",
840 "\nBased on the difference in skewness we would assume \ncontamination of
841 healthy group by false negative \nsubjects from the",
842 ifelse(skew.higher=="test-positive","lower","upper"),
843 "tail of the distribution \nof infected individuals, which would lead to
844 overall",
845 ifelse(skew.higher=="test-positive","\ndecrease","\nincrease"),
846 "of test-negative group mean.))
847 ,
848 ifelse(skew.sig==FALSE,"",
849 paste(ifelse(set.higher==skew.guess.higher,
850 paste("\n\nThe skewness analysis brings further support to the hypothesis
851 \nof",
852 ifelse(set.higher,"higher","lower"),
853 "mean in non-contaminated group of healthy \nindividuals, which was used in
854 current permutation test \nfor contaminated data.\n"),
855 paste("\n\nThe skewness analysis, however, does not support the hypothesis
856 \nof",
857 ifelse(set.higher,"higher","lower"),
858 "mean in non-contaminated group of healthy \nindividuals, which was used in
859 current permutation test \nfor contaminated data. Proceed with caution.\n")
860 )))
861 )
862
863 skewness<-rbind(skewness[1], "", skew.p)
864
865 rownames(skewness)[c(3,4)]<-c("", "p-value")
866
867 skewness[c(1,2,4),1]<-format(round(as.numeric(skewness[c(1,2,4),1]),3))
868
869 mean.dist.perm<-rowMeans(perm.dist)
```

```
870
871 for(i in 1:Nperc){
872   p.vals.perm[i]<-sum(dist.reals[i]<perm.dist[i,])/runs
873 }
874
875 mean.healthy<-format(round(mean.healthy,2))
876 mean.infected<-format(round(mean.infected,2))
877 sd.healthy<-format(round(sd.healthy,2))
878 sd.infected<-format(round(sd.infected,2))
879 cohen<-format(round(cohen,2))
880 dist.reals<-format(round(dist.reals,2))
881
882 mean.dist.perm<-format(round(mean.dist.perm,2))
883 p.vals.perm<-format(round(p.vals.perm,3))
884
885 res.table<-
886 rbind(contamination,mean.healthy,mean.infected,dist.reals,mean.dist.perm,sd
887 .healthy,sd.infected,N.healthy,N.infected,cohen,p.vals.perm)
888 res.table<-as.data.frame(res.table)
889
890 colnames(res.table)<-NULL
891 rownames(res.table)<-c("contamination","non-infeceted mean","infected
892 mean","mean difference","expected difference","non-infeceted SD","infected
893 SD","non-infeceted N","infected N","Cohen's d","p-value")
894
895 final.message<-paste("\nTwo-tailed permutation test for contaminated data
896 was executed. \nIt was assumed that",
897 ifelse(set.higher==T,"healthy","infected"),
898 "individuals have on average higher \ntrait value if all false negative
899 individuals are relocated correctly.")
900
901 cat("\nSample characteristics:\n")
902 print(orig.means)
```

```
903   cat(paste("\n",higher.report,"\n\n",sep=""))
904   cat(set.report1)
905
906   if(skewness.analysis==T){
907     cat("\n\nSkewness report:\n")
908     print(skewness)
909     cat("\n")
910     cat(skew.message)
911   }
912
913   cat("\n\nPermutation test for contaminated data:\n")
914   cat(paste(set.report2,"\n\n",sep=""))
915   cat(which.test)
916   cat(paste("\n",runs,"sample permutations were performed.\n"))
917   print(res.table)
918   cat(paste(final.message,"\n\n",sep=""))
919
920   results<-
921   list(orig.means,higher.report,set.report1,skewness,skew.message,set.report2
922   ,which.test,res.table,final.message)
923
924   return(invisible(results))
925 }
926
```


927 **Supplement 2**

928 **Skewness analysis**

```
929 #####
930 #####hit ctrl+a and ctrl+r to install the function####
931 #####
932
933 #This script contains skewness_comparison() function
934 #skewness_comparison(trait,identification,percentages=seq(0,50,1),higher.healthy=
935 (mean(trait[identification==F])>mean(trait[identification==T])),runs=
936 10000)
937
938 #Arguments are described below.
939 #It is necessary to define function calculating skewness index first:
940
941 ##Function that returns Fisher-Pearson coefficient of skewness.
942 ##Input is a vector of numerical values.
943
944 FPskewness<-function(x){
945   return((sum((x-mean(x))^3)/length(x))/((sqrt(sum((x-
946   mean(x))^2)/length(x))^3))
947 }
948
949 ###Skewness comparison
950 ###Function that reports how relocation of seronegative individuals changes
951 the skewness of the distribution in both seronegative and seropositive
952 groups.
953 ###trait - Numerical vector of trait values
954 ###identification - Logical vector of assumed presence (T) or absence (F)
955 of infection
956 ###percentages - Numerical vector of percentages of false negative amongst
957 negative subjects (contamination levels) for which the permutation test for
958 contaminated data will be run
959 ###higher healthy - Logical. Indicates whether we assume the healthy
960 individuals to show higher (T) or lower (F) trait values. When not
```

```
961 specified, the script assumes this relationship based of group means with
962 no hypothesised contamination.

963 #####It allows us to use the difference between the groups (not in
964 absolute values) in permutation test. In this scenario the seropositive
965 group mean is substracted from seronegative group mean and the one-tailed
966 permutation test is conducted accordingly.

967 ###runs - Number of resamplings used in permutation test

968

969 skewness_comparison<-
970 function(trait,identification,percentages=seq(0,50,1),higher.healthy=(mean(
971 trait[identification==F])>mean(trait[identification==T])),runs=10000){

972

973 if(length(trait)!=length(identification)){

974 stop("The vectors of trait values and infection indication are of different
975 lengths.")

976 }

977

978 higher<-(mean(trait[identification==F])>mean(trait[identification==T]))
979 set.higher<-higher.healthy

980

981 orig.means<-tapply(trait,identification,mean)
982 orig.means<-data.frame(orig.means)

983

984 names(orig.means)<-"Original mean values"

985 rownames(orig.means)[which(rownames(orig.means)=="FALSE")]<-"Identified as
986 healthy"

987 rownames(orig.means)[which(rownames(orig.means)=="TRUE")]<-"Identified as
988 infected"

989

990 higher.report<-ifelse(higher==T,

991 "In original sample, individuals identified as healthy showed higher
992 \naverage trait value.",

993 "In original sample, individuals identified as infected showed higher
994 \naverage trait value."

995 )
```

```
996
997 concord<-ifelse(higher==set.higher,"Consequently","Despite that,")
998
999 set.report1<-paste(concord,ifelse(set.higher==T,
1000 "healthy individuals were hypothesised to have higher \naverage trait value
1001 in a contamination-free sample. \n",
1002 "infected individuals were hypothesised to have higher \naverage trait
1003 value in a contamination-free sample. \n"
1004 ))
1005
1006
1007 run.report<-paste("Two-tailed permutation test of skewness difference \non
1008 ",runs," runs was executed on each contamination level.\n\n",sep="")
1009
1010 set.report2<-paste("\nFor each contamination level respective proportion of
1011 seronegative \nindividuals with",
1012 ifelse(set.higher==T,"lowest","highest"),
1013 "trait value was relabeled as seropositive \nand the difference between the
1014 group skewness was measured. \nReferential skewness differences from
1015 permutation runs were based \non random non-contaminated sample with group
1016 sizes corresponding \nto respective contamination levels.\n\n"
1017 )
1018
1019 trait<-c(trait[identification==F],trait[identification==T])
1020 infected<-sort(identification)
1021
1022 count.healthy<-sum(!identification)
1023 count.infected<-sum(identification)
1024
1025 Nperc<-length(percentages)
1026
1027 vector.ident<-list()
1028
```

```
1029 for(i in 1:Nperc){
1030   reassign<-round(count.healthy*(percentages[i]/100))
1031   identification<-c(rep(F,count.healthy-
1032     reassign),rep(T,count.infected+reassign))
1033   vector.ident[[i]]<-identification
1034 }
1035
1036 #Sorts healthy individuals to indicate possible false-negatives
1037 trait2<-
1038 c(sort(trait[infected==F],decreasing=higher.healthy),trait[infected==T])
1039
1040 skewness.healthy<-1:Nperc
1041 names(skewness.healthy)<-paste(as.character(percentages), "%")
1042 skewness.infected<-skewness.healthy
1043
1044 contamination<-paste(as.character(percentages), "%")
1045 names(contamination)<-paste(as.character(percentages), "%")
1046
1047 #Compute group means in non-permuted sample
1048 for(i in 1:Nperc){
1049   skewness.healthy[i]<-FPskewness(trait2[vector.ident[[i]]==F])
1050   skewness.infected[i]<-FPskewness(trait2[vector.ident[[i]]==T])
1051 }
1052
1053 skew.diff<-abs(skewness.healthy-skewness.infected)
1054
1055 #Permutation test of skewness difference
1056
1057 skew.diff.perm<-array(NA,dim=c(Nperc,runs))
1058
1059 for(run in 1:runs){
```

```
1060 trait2<-sample(trait)
1061
1062 for(i in 1:Nperc){
1063 skew.diff.perm[i,run]<-abs(FPskewness(trait2[vector.ident[[i]]==F]) -
1064 FPskewness(trait2[vector.ident[[i]]==T]))
1065 }
1066 }
1067
1068 p.vals.skew<-NA
1069
1070 for(i in 1:Nperc){
1071 p.vals.skew[i]<-sum(skew.diff[i]<skew.diff.perm[i,])/runs
1072 }
1073
1074 guess.perc<-percentages[which(skew.diff==min(skew.diff))]
1075
1076 possible<-percentages[p.vals.skew>0.05]
1077
1078 if(length(possible)==0){
1079 ok.report<-paste("\nThe difference in Fisher-Pearson coefficients of
1080 skewness between \ngroups was significant at all investigated contamination
1081 levels.\nThis may be caused by extreme proportion of false negative
1082 \nindividuals, insufficient number of relocated fractions, \ndifferent
1083 shapes of distributions of healthy and infeceted \nindividuals, or, most
1084 likely, by wrong setting of higher group \nmean in non-contaminated sample.
1085 \nTry running this comparison with parameter higher.healthy
1086 =",ifelse(higher.healthy==TRUE,"FALSE","TRUE"),"\n\n")
1087 }else{
1088 if(min(possible)==max(possible)){
1089 ok.report<-paste("\nThe difference in Fisher-Pearson coefficients of
1090 skewness between \ngroups was not significant at",
1091 min(possible),
1092 "% of relocated individuals.\n\n")
1093 }else{
```

```
1094 ok.report<-paste("\nThe difference in Fisher-Pearson coefficients of
1095 skewness between \ngroups was not significant between",
1096 min(possible) ,
1097 "% and",
1098 max(possible) ,
1099 "% of relocated individuals.\n\n")
1100 }
1101 }
1102
1103 best.guess<-paste("The difference between group skewness was smallest
1104 \nwhen",
1105 guess.perc,
1106 "% of seronegative individuals with",
1107 ifelse(higher.healthy, "lowest", "highest"),
1108 "trait value \nwas relocated to seropositive group.\n\n")
1109
1110
1111 skewness.healthy<-format(round(skewness.healthy,2))
1112 skewness.infected<-format(round(skewness.infected,2))
1113 p.vals.skew<-format(round(p.vals.skew,3))
1114
1115 skewness.res<-
1116 rbind(contamination,skewness.healthy,skewness.infected,"",p.vals.skew)
1117
1118 skewness.res<-as.data.frame(skewness.res)
1119
1120 colnames(skewness.res)<-NULL
1121 rownames(skewness.res)<-c("contamination","skewness healthy","skewness
1122 infected","", "p-value")
1123
1124
1125 cat("\nSample characteristics:\n")
```

```
1126 print(orig.means)
1127 cat(paste("\n",higher.report,"\n\n",sep=""))
1128 cat(set.report1)
1129 cat(set.report2)
1130
1131 cat(run.report)
1132
1133 cat(paste("Skewness comparison:", "\n", sep=""))
1134
1135 print(skewness.res)
1136
1137 cat(ok.report)
1138 cat(best.guess)
1139
1140 results<-list(orig.means,higher.report,set.report1,skewness.res)
1141
1142 return(invisible(results))
1143 }
1144
```

1145 **Supplement 3**

1146 **Example**

1147 #####

1148 ###Exemplar runs###

1149 #####

1150

1151 #Both functions contamination_perm_test() and skewness_comparison() must be
1152 installed, run respective scripts

1153

1154 #You can generate data with known proportion of false negative individuals
1155 with this script and try permutation test for contaminated data on them.

1156

1157 count<-1000 #sets sample size

1158 inf.prop<-0.5 #sets proportion of seropositive individuals

1159 fixed.effect<-(-3) #sets the effect of infection on simulated trait

1160 healthy.average<-101.5 #sets the average trait value in noninfected group

1161 sd<-12 #sets the standard deviation of within group

1162

1163 false.negatives<-5 #sets proportion of false negative individuals

1164

1165 ###Computes counts in respektive groups using set properties

1166 count.infected<-round(inf.prop*count)

1167 count.healthy<-count-count.infected

1168

1169 ###Creates a variables that indicates infection

1170 infected<-c(rep(F,count.healthy),rep(T,count.infected))

1171

1172 #Calculates how many false negative individuals will be in the seronegative
1173 group

1174 reassign<-round(count.healthy*(false.negatives/100))

1175 #Creates a vector of actual iinfection


```
1176 really.infected<-c(rep(F,count.healthy-
1177 reassign),rep(T,count.infected+reassign))
1178
1179 #Generates the population (all healthy individuals)
1180 trait<-rnorm(count,healthy.average,sd)
1181
1182 #modifies really infected individuals
1183 trait[(sum(!really.infected)+1):count]<-
1184 trait[(sum(!really.infected)+1):count]+fixed.effect
1185
1186 #sorts really infeceted individuals such that most changed ones are close
1187 in the vector to healthy ones i.e. are marked as test-negative
1188
1188 trait<-
1189 c(trait[really.infected==F],sort(trait[really.infected==T],decreasing=(sign
1190 (fixed.effect)==1)))
1191
1192 #scrambles the vectors, along the same random vector
1193
1193 scramble<-sample(1:count)
1194
1195
1195 infected<-infected[scramble]
1196
1196 trait<-trait[scramble]
1197
1198 #####
1199 ###Trial data are ready###
1200 #####
1201
1202 #Executes permutation test for contaminated data with default argument
1203 values
1204
1204 contamination_perm_test(trait,infected)
1205
1206 #Executes permutation test for contaminated data and skewness analysis
1207
1207 contamination_perm_test(trait,infected,skewness.analysis=T)
1208
```

```
1209 #Executes permutation test for contaminated data, hypothesised difference is
1210 specified by hand. This comes useful when you have a reason to suspect
1211 paradoxical switch in group means.
1212 contamination_perm_test(trait,infected,higher.healthy=F)
1213
1214 #Executes permutation test on only 100 permutation runs per test - it is
1215 quicker, but less accurate
1216 contamination_perm_test(trait,infected,runs=100)
1217
1218 #Executes two tailed permutation test for contaminated dat
1219 contamination_perm_test(trait,infected,two.tailed=T)
1220
1221 #Executes skewness comparison to estimate proportion and distribution tail
1222 of possible contamination prior to the test
1223 skewness_comparison(trait,infected)
1224
1225 #Executes permutation test for contaminated data, levels of contamination
1226 are specified by hand
1227 contamination_perm_test(trait,infected,percentages=seq(1,15,1))
```