A comparison of common metrics used to quantify the effectiveness of conservation interventions

Igor Khorozyan [Coresp. 1]

[1] Workgroup on Endangered Species, J.F. Blumenbach Institute of Zoology and Anthropology, Georg-August Universität Göttingen, Göttingen, Germany

Corresponding Author: Igor Khorozyan
Email address: igor.khorozyan@biologie.uni-goettingen.de

Background. Evidence-based conservation is urgently needed to identify, apply and promote effective interventions for mitigation of threats and recovery of the natural environment. Estimation of intervention effectiveness is subject to robust study design and statistical analysis, and much progress is documented in these fields. In contrast, little is understood about the accuracy and biases (underestimation and overestimation) of different effectiveness metrics and how they are affected by sample size.

Methods. In this study, a dataset (n = 500 cases) consisting of random, positive, integer numbers was simulated to produce frequency input data for the 2×2 contingency table. For each case, three metrics of the relative risk, odds ratio and the magnitude of change were calculated, their disparity was estimated and the characteristics of treatment (with intervention) and control (without intervention) samples significantly affecting this disparity were studied by means of linear regression. Further, four case studies from different conservation interventions are provided to support the results.

Results. The study has shown that the relative risk and the magnitude of change produce identical estimates of intervention effectiveness only when treatment and control samples are equal, or when the number of target outcomes (e.g., number of livestock killed by predators) in treatment sample reaches zero. In other situations, the magnitude of change gives overestimates or underestimates, depending on relationships between treatment and control sample sizes. The table summarizing the conditions of equalities and biases between these two metrics is provided. These conditions are valid for both reduction-aimed interventions reducing negative target outcomes (e.g., livestock protection to reduce livestock losses to predators) and for addition-aimed interventions increasing positive target outcomes (e.g., establishment of protected areas to increase species presence). No significant effects on the odds ratio were found.

Conclusion. Researchers should set equal treatment and control sample sizes so that to produce identical estimates of intervention effectiveness by the relative risk and the magnitude of change. Otherwise, these estimates are biased if produced by the magnitude of change and the relative risk should be used instead. As setting equal treatment and control samples can be impractical, I encourage researchers and practitioners to use the relative risk in estimation of intervention effectiveness. This will not take additional efforts as both metrics are calculated from the same contingency table. Treatment and control sample sizes, along with their sub-samples affected and not by an intervention, should be explicitly reported by researchers to allow independent evaluation of intervention effectiveness. This approach can help obtain more accurate information on intervention effectiveness for making better decisions in conservation actions.

1   **A comparison of common metrics used to quantify the**
2   **effectiveness of conservation interventions**
3
4
5   Igor Khorozyan
6
7   Workgroup on Endangered Species, J.F. Blumenbach Institute of Zoology and Anthropology,
8   Georg-August-Universität Göttingen, Bürgerstr. 50, Göttingen 37073, Germany. E-mail:
9   igor.khorozyan@biologie.uni-goettingen.de
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35

Abstract

Background. Evidence-based conservation is urgently needed to identify, apply and promote effective interventions for mitigation of threats and recovery of the natural environment. Estimation of intervention effectiveness is subject to robust study design and statistical analysis, and much progress is documented in these fields. In contrast, little is understood about the accuracy and biases (underestimation and overestimation) of different effectiveness metrics and how they are affected by sample size.

Methods. In this study, a dataset (n = 500 cases) consisting of random, positive, integer numbers was simulated to produce frequency input data for the 2×2 contingency table. For each case, three metrics of the relative risk, odds ratio and the magnitude of change were calculated, their disparity was estimated and the characteristics of treatment (with intervention) and control (without intervention) samples significantly affecting this disparity were studied by means of linear regression. Further, four case studies from different conservation interventions are provided to support the results.

Results. The study has shown that the relative risk and the magnitude of change produce identical estimates of intervention effectiveness only when treatment and control samples are equal, or when the number of target outcomes (e.g., number of livestock killed by predators) in treatment sample reaches zero. In other situations, the magnitude of change gives overestimates or underestimates, depending on relationships between treatment and control sample sizes. The table summarizing the conditions of equalities and biases between these two metrics is provided. These conditions are valid for both reduction-aimed interventions reducing negative target outcomes (e.g., livestock protection to reduce livestock losses to predators) and for addition-aimed interventions increasing positive target outcomes (e.g., establishment of protected areas to increase species presence). No significant effects on the odds ratio were found.

Conclusion. Researchers should set equal treatment and control sample sizes so that to produce identical estimates of intervention effectiveness by the relative risk and the magnitude of change. Otherwise, these estimates are biased if produced by the magnitude of change and the relative risk should be used instead. As setting equal treatment and control samples can be impractical, I encourage researchers and practitioners to use the relative risk in estimation of intervention effectiveness. This will not take additional efforts as both metrics are calculated from the same contingency table. Treatment and control sample sizes, along with their sub-samples affected and not by an intervention, should be explicitly reported by researchers to allow independent evaluation of intervention effectiveness. This approach can help obtain more accurate information on intervention effectiveness for making better decisions in conservation actions.

Introduction

Coping with globally widespread human pressures on the environment demands the urgent promotion of evidence-based conservation, which is the use of best available evidence to foster informed decision-making in practical conservation (*Pullin et al., 2004*; *Sutherland et al., 2004*; *Sutherland & Wordley, 2017*; *Altringham, Berthinussen & Wordley, 2020*). This concept ensues from, and follows the principles of, the effectiveness revolution which began in the 1970s in medicine (*Pullin & Knight, 2001*). In other words, evidence-based conservation strives to collect, analyze, interpret, synthesize and distribute information on the effectiveness of interventions applied to mitigate threats and improve the status of (semi-)natural ecosystems and their components (*Altringham, Berthinussen & Wordley, 2020*; *Westgate et al., 2018*). Evidence is subject to rigorous scientific assessments and it should be widely published to stimulate replications and reduce the research-implementation gap (*Toomey, Knight & Barlow, 2017*; *Burivalova et al., 2019*). Comparisons of several interventions in the same setting are very useful to select the most effective interventions for further applications (*Smith et al., 2014*; *Schmidt et al., 2019*). Active development of open-access web resources (e.g., Conservation Evidence – www.conservationevidence.com and the Collaboration for Environmental Evidence – www.environmentalevidence.org), professional meetings, analytical tools and outreach awareness-raising is aimed at the wider use of evidence in conservation actions (*Sutherland & Wordley, 2017*; *Westgate et al., 2018*; *Sutherland et al., 2019*; *Pullin et al., 2020*).

Deeper understanding of the importance of intervention effectiveness leads to more scrutiny over the issues of study design and statistical analysis. It is generally agreed that experiments must be controlled and case-control studies, such as comparisons before and after the intervention and comparisons with and without the intervention, are less reliable because they do not have properly matched intervention-free controls to compare with (*Burivalova et al., 2019*; *Christie et al., 2019*). Arguably, the most reliable study designs are before-after-control-impact (BACI), control trials and cross-over, especially when study objects are randomized and experimental procedures are blinded to minimize biases (*Adams, Barnes & Pressey, 2019*; *Christie et al., 2019*; *Treves et al., 2019*). A number of metrics have been used to quantify the effectiveness of interventions, including the relative risk (*Eklund et al., 2017*; *Khorozyan & Waltert, 2019a*; *Khorozyan & Waltert, 2019b*; *Bruns, Waltert & Khorozyan, 2020*; *Khorozyan et al., 2020*), odds ratio (*Knarrum et al., 2006*; *Woodroffe et al., 2007*; *Wielgus & Peebles, 2014*), magnitude of change (*Jones & Schmitz, 2009*; *Green et al., 2013*; *Miller et al., 2016*) and Hedges' d (*van Eeden et al., 2017*).

Despite progress in developing the methodology of evidence-based conservation, very little understanding has been obtained regarding differences in effectiveness estimates produced by different metrics. Scientists usually use a single metric per study and it remains unclear whether their estimates are sufficiently accurate or produce biases (underestimation and overestimation) for a variety of reasons. The relative risk and the odds ratio are very similar when events are rare (*Stare & Maucort-Boulch, 2016*), e.g. when predators kill livestock and the effectiveness of

116 livestock protection interventions is evaluated (*Khorozyan & Waltert, 2019a*; *Bruns, Waltert &*
117 *Khorozyan, 2020*). *Van Eeden et al. (2018)* compiled three independent meta-analyses, each of
118 which used a different metric (relative risk, magnitude of change and Hedges' d), to assemble
119 current information on the effectiveness of livestock protection interventions against predators,
120 but did not compare these metrics from the same case studies.
121   It is of particular practical importance and scientific interest to compare the metrics
122 calculated from the same standardized matrix produced from the same study. Among the most
123 common matrices is the 2×2 contingency table which summarizes the distribution of
124 dichotomous outcomes, such as species A present and species A absent, across the treatment
125 (with intervention) and control (without intervention) samples (*Stare & Maucort-Boulch, 2016*).
126 The contingency table approach can be used to calculate the relative risk, odds ratio and the
127 magnitude of change with no need for additional information. The effects of treatment and
128 control samples and their relationships on metrics are key to identify possibilities for accurate
129 and unbiased estimation of effectiveness. Larger samples improve the precision of effectiveness
130 estimates (*Smith et al., 2014*; *Christie et al., 2019*), but this is not sufficient and the relationships
131 between treatment and control samples may uncover new effects worthy of further
132 investigations. As a researcher is responsible for setting treatment and control sample sizes
133 before the study, this knowledge may provide useful rules-of-thumb for study design,
134 implementation and data analysis.
135   In this study, I used a simulated dataset to calculate the relative risk, odds ratio and
136 magnitude of change from the 2×2 contingency table, estimate their disparity, and find sample
137 characteristics significantly affecting this disparity. I did not consider Hedges' d as this metric
138 requires additional quantitative information (mean and standard deviation) from treatment and
139 control samples, which is not available from contingency tables (*Nakagawa & Cuthill, 2007*).
140 Four case studies from different conservation interventions are provided and described in order
141 to illustrate and support the results.
142
Materials & Methods
144
Framework
146
147   The 2×2 contingency table of the numbers of target and alternative outcomes in treatment
148 and control samples was a basis of this study (Fig. 1). The target outcomes are the outcomes that
149 interventions strive to change, otherwise they are designated as the alternative outcomes. As
150 interventions aim to reduce negative target outcomes or add positive target outcomes, they were
151 provisionally named as reduction-aimed interventions and addition-aimed interventions. An
152 example of reduction-aimed intervention is livestock protection that aims to reduce or eliminate
153 livestock losses to predators (*Eklund et al., 2017*) and an example of addition-aimed intervention
154 is the establishment of protected areas to increase species presence (*da Silva et al., 2018*). In
155 these two examples, the target outcomes are the number of lost livestock and the number of

156 sampling units (e.g., grids or transects) with species presence. The alternative outcomes are the
157 number of surviving livestock and the number of sampling units with species absence. The
158 treatment samples are those affected by interventions and the control samples, also known as
159 counterfactuals, are not affected.
160   The purpose of data analysis from the contingency table is to determine the effectiveness of
161 interventions, i.e. whether they significantly decrease the target outcomes for reduction-aimed
162 interventions or significantly increase the target outcomes for addition-aimed interventions. In
163 Fig. 1, N_{t1} stands for the number of target outcomes in treatment sample, N_{t2} for the number of
164 alternative outcomes in treatment sample, N_{c1} for the number of target outcomes in control
165 sample and N_{c2} for the number of alternative outcomes in control sample. The treatment sample
166 size is $N_t = N_{t1} + N_{t2}$ and the control sample size is $N_c = N_{c1} + N_{c2}$.

167

168 Data

169

170   I simulated a dataset of N_{t1}, N_{t2}, N_{c1} and N_{c2} which consisted of random, positive, integer
171 numbers and included 500 independent cases (Data S1). For this, I used the
172 RANDBETWEEN(1,1000) function in MS Excel to select random integers between 1 and 1000.
173 The numbers N_t and N_c were calculated accordingly. It was assumed that each case represents an
174 individual and independent study.
175   Intervention effectiveness was estimated for each case in comparable metrics of the relative
176 risk (RR%), odds ratio (OR%) and the magnitude of change (D%), all presented in percentages
177 of outcome change (decrease, no change, increase). Equations 1-3 use the notations described
178 above. Abbreviations RR, OR and D stand for proportions not transformed to percentages.

179

180

181       $$RR\% = (RR - 1) \times 100 = \left(\frac{N_{t1}/N_t}{N_{c1}/N_c} - 1 \right) \times 100 \qquad (1)$$

182

183

184       $$OR\% = (OR - 1) \times 100 = \left(\frac{N_{t1}/N_{t2}}{N_{c1}/N_{c2}} - 1 \right) \times 100 \qquad (2)$$

185

186

187       $$D\% = \left(\frac{N_{t1}}{N_{c1}} - 1 \right) \times 100 \qquad (3)$$

188

189

190   The term (RR - 1) was taken instead of (1 - RR) (*Khorozyan & Waltert, 2019b*; *Bruns,*
191 *Waltert & Khorozyan, 2020*) to make RR%, OR% and D% move in the same direction: their
192 negative values indicate a decrease in target outcome, 0 indicate no change and their positive
193 values indicate an increase in target outcome. For the same reason, we inverted the nominator
194 provided in the original equation of D% (*Jones & Schmitz, 2009*; *Green et al., 2013*; *Miller et*

195 *al., 2016*). The minimum values of RR%, OR% and D% are -100%, meaning that the target
196 outcome decreases by 100% when $N_{t1} = 0$.
197   The measures of disparity between RR%, OR% and D% were the absolute differences
198 |RR% - D%|, |RR% - OR%| and |OR% - D%|, which I called the disparity distance. The disparity
199 distance can be caused by both underestimation and overestimation by any of these metrics,
200 therefore it was used as an absolute value for standardization.
201
202 Data analysis
203
204   As the three effectiveness metrics depend on outcome proportions, the effects of the
205 following outcome ratios on the disparity distance were measured: N_{t1}/N_{t2}, N_{c1}/N_{c2}, N_{t1}/N_{c1},
206 N_{t2}/N_{c2}, N_{t1}/N_t, N_{t2}/N_t, N_{c1}/N_c, N_{c2}/N_c and N_t/N_c. The relationships between each of these ratios
207 and the disparity distance were checked for statistical significance and the effect size (coefficient
208 of determination r^2) using scatterplots and simple linear regression (*Guthery & Bingham, 2007*;
209 *McKillup, 2012*). Outliers were defined as the points with Cook's distance > 1 and removed for
210 further re-analysis. Precision was measured as the 95% confidence intervals of the line and the
211 non-zero slope β_1, and accuracy as the root mean square error (RMSE) which was normalized
212 (nRMSE) by taking its percentage from the range $y_{max} - y_{min}$ (*Wan et al., 2010*). The effects of
213 outcome ratios on the disparity distance were checked by significance of the $F_{1,n-2}$ statistic at p <
214 0.05 (*McKillup, 2012*).
215   After significant predictors were found among the outcome ratios, the regression lines of
216 these predictors were built against RR%, OR% and D% to visualize the disparity distance, with
217 further estimation of precision, accuracy and effect size as described above. The significance of
218 difference between the slopes of RR%, OR% and D% lines was estimated using the interaction
219 terms in ANCOVA (*McKillup, 2012*). To determine underestimation and overestimation between
220 the metrics, significant outcome ratios were split into two almost equal groups, < 1 and >1, and
221 the pairs RR% vs. OR%, OR% vs. D% and RR% vs. D% were compared between these two
222 groups by Mann-Whitney test. Chi-square (χ^2) test was applied to compare the frequencies of
223 RR%, OR% and D%. The program IBM SPSS 26.0 was used for statistical analysis throughout
224 the study.
225
226 Case studies
227
228   To demonstrate wide applicability of study results, I attempted to use four case studies
229 focused on mammals, birds, reptiles and landscape restoration, dealing with reduction-aimed and
230 addition-aimed interventions. I used my library (*Knarrum et al., 2006*), Conservation Evidence
231 (*Newell, 2019*), and Google Scholar (http://scholar.google.com, search words "turtle 2x2
232 contingency table", *Sampaio, 2018*; "restoration 2x2 contingency table", *Andruk, 2014*) to find
233 the literature. These publications contain data produced from controlled experiments. Original

234   data of these studies were used to produce contingency tables and calculate the metrics of RR%,
235   OR% and D% for comparison.
236
Results
238
239   Disparity distances between OR% and the other two metrics were significantly affected
240   only by the odds of target outcomes in treatment sample (N_{t1}/N_{t2}), which is natural as OR% is
241   calculated from these odds (see equation 2 above). Therefore, only RR% and D% were
242   considered further in this study. The only outcome ratio significantly affecting the disparity
243   distance between RR% and D% was the ratio of the number of target outcomes in treatment
244   sample (N_{t1}) to that in control sample (N_{c1}), i.e. N_{t1}/N_{c1} (Fig. 2). This effect was significant ($F_{1,480}$
245   = 1194.49, $p < 0.001$), precise and positive ($\beta_1 = 51.402 \pm 1.487$, 95% CI = 48.480-54.324) and
246   strong ($r^2 = 0.713$), but not accurate (nRMSE = 575.63%). As D% is calculated from N_{t1}/N_{c1}, its
247   line was statistically perfect ($r^2 = 1$). The effect of N_{t1}/N_{c1} on RR% was also significant ($F_{1,489} =$
248   2081.50), precise and positive ($\beta_1 = 52.571 \pm 1.152$, 95% CI = 50.307-54.835), strong ($r^2 =$
249   0.810) and accurate (nRMSE = 3.33%) (Fig. 2). The slopes of D% and RR%, which embraced
250   the disparity distance within, were statistically different (ANCOVA: $F_{1,982} = 2036.80$, $p < 0.001$).
251   When N_{t1}/N_{c1} was less than 1, there was a significantly higher number of cases with D% <
252   RR% (n = 179) than with D% > RR% (n = 75), and this difference was significant ($\chi^2 = 42.583$,
253    $p < 0.001$). The estimate D% was significantly lower than RR% (Mann-Whitney U = 25913.5, p
254   < 0.001). These inequalities depended on the ratio of treatment sample size (N_t) to control
255   sample size (N_c): |D%| < |RR%| when N_t/N_c was higher than 1 and |D%| > |RR%| when N_t/N_c
256   was less than 1. When N_t/N_c is equal to 1, D% = RR%. The absolute estimates are provided
257   because D% and RR% can be both positive and negative. The general scheme of
258   underestimations, overestimations and equalities of D% and RR% is shown in Table 1.
259   When N_{t1}/N_{c1} was higher than 1, there was a significantly higher number of cases with D%
260   > RR% (n = 172) than with D% < RR% (n = 64), and this difference was significant ($\chi^2 =$
261   49.424, $p < 0.001$). The estimate D% was significantly higher than RR% (Mann-Whitney U =
262   18779.0, $p < 0.001$). Like in the previous scenario, these relationships depended on N_t/N_c: |D%|
263   > |RR%| when N_t/N_c was higher than 1 and |D%| < |RR%| when N_t/N_c was less than 1. Again,
264   when N_t/N_c is equal to 1, D% = RR% (Table 1).
265   When N_{t1}/N_{c1} is equal to 1, D% indicates no effect of interventions regardless of
266   relationships between N_t and N_c whereas RR% indicates a decrease or increase of target
267   outcomes depending on sample sizes. When $N_t = N_c$, both D% and RR% estimate no effect of
268   interventions (Table 1).
269
270
271
272

273 Case studies

274

275 Reduction-aimed interventions. Case 1: Use of bell collars to deter brown bears (*Ursus arctos*)
276 from ewes in Norway

277

278   Bell collars are often used to easily locate grazing livestock. However, they also may serve
279 as acoustic deterrents from predators, thus reducing livestock mortality from predator attacks and
280 relieving tensions between local people and conservationists. To test this, *Knarrum et al. (2006)*
281 compared depredation rates among the ewes with and without bell collars caused by brown bear
282 attacks. The contingency table is shown in Table 2. In this sample, $N_{t1} > N_{c1}$ and $N_t < N_c$,
283 therefore the absolute estimate of D% should be lower than RR% (Table 1). Our results show
284 that bell collars increased ewe mortality from bear attacks by D% = 42.8% and RR% = 155.6%
285 instead of decreasing it, possibly because bears learned to associate the sound of bells with food.
286 So, bell collars were ineffective and even counter-productive in protecting ewes from bears.

287

288 Reduction-aimed interventions. Case 2: Use of metal nets to protect green turtle (Chelonia
289 mydas) nests from predation in Guinea-Bissau

290

291   The green turtle is globally endangered due to uncontrolled harvesting of animals and eggs,
292 and the main nesting grounds in undisturbed sites need protection from predators. In this study,
293 randomly selected turtle nests were covered by 1 m^2 metal nets to prevent access to Nile
294 monitors (*Varanus niloticus*) and compared with open control nests in terms of monitor
295 predation (*Sampaio, 2018*; Table 3). As $N_{t1} < N_{c1}$ and $N_t < N_c$ in this sample, the absolute D% is
296 to be higher than RR% (Table 1). The result is D% = -84.2% and RR% = -61.1%, confirming
297 that metal nets were effective in reducing nest predation.

298

299 Addition-aimed interventions. Case 3: Use of nest forms to increase the occupancy of nest boxes
300 by common swifts (*Apus apus*) in the UK

301

302   As swifts are endangered in the UK, principally due to the lack of suitable nesting sites,
303 nest boxes are set up to recover the population. Yet, the optimal design of nest boxes is
304 insufficiently known and placing a nest form inside nest boxes is a promising method to increase
305 the occupancy of nest boxes by swifts. In this study, the numbers of occupied nest boxes were
306 compared between the boxes with and without nesting forms (*Newell, 2019*; Table 4). Here, N_{t1}
307 > N_{c1} and $N_t < N_c$, meaning the absolute D% being lower than RR% (Table 1). Calculations
308 show that nesting forms increased the occupancy of nesting boxes by swifts by D% = 150.0%
309 and RR% = 222.6%, and their use is effective for swift conservation.

310

311

312   Addition-aimed interventions. Case 4: Use of prescribed fire to recover woodland in the USA
313
314   Prescribed fire is a common intervention used to recover natural vegetation of arid
315   landscapes. *Andruk (2014)* tested how the fire increased the number of sprouts of hardwood
316   plants such as sugarberry (*Celtis laevigata*), black cherry (*Prunus serotina*) and others in
317   woodland of central Texas. In this study design, the number of sprouts post-treatment served as a
318   target outcome and the number of sprouts pre-treatment was an alternative outcome (Table 5).
319   With $N_{t1} > N_{c1}$ and $N_t > N_c$, the absolute D% should be higher than RR% (Table 1). The result is
320   that prescribed fire increased the number of hardwood sprouts by D% = 340.1% and RR% =
321   17.2%.
322

Discussion

324
325   This study has shown that the relative risk and the magnitude of change produce identical
326   estimates of intervention effectiveness only when treatment and control samples are equal, or
327   when the number of target outcomes in treatment sample is zero (Table 1). In other situations,
328   the magnitude of change gives overestimates or underestimates, depending on relationships
329   between treatment and control sample sizes. These patterns were demonstrated by means of
330   theoretical simulations and practical case studies, for both reduction-aimed and addition-aimed
331   conservation interventions. Overestimation takes place when the relationships between treatment
332   and control sample sizes, and between the numbers of target outcomes in treatment and control
333   samples, are in the same direction – i.e. both are lower, or both are higher, in treatment than in
334   control samples (Table 1). When these relationships go in opposite directions, the magnitude of
335   change gives underestimation of effectiveness. The higher the dissimilarity between treatment
336   and control sample sizes, the higher the magnitude of overestimation or underestimation, which
337   is best illustrated by case studies 1 and 4 (*Knarrum et al., 2006*; *Andruk, 2014*). Disparity
338   between the two metrics further increases with the increasing ratio of the number of target
339   outcomes in the treatment sample to that in the control sample (Fig. 2).
340   Information about the accuracy of different effectiveness metrics under variable
341   relationships between treatment and control sample sizes is practically important because the
342   determination of sample size is the researcher's prime task to design a good study (*Smith et al.,
343   2014*; *Christie et al., 2019*). When samples are subject to study design, as in the cases with bell-
344   collared and uncollared ewes (case study 1; *Knarrum et al., 2006*), net-covered and uncovered
345   turtle nests (case study 2; *Sampaio, 2018*) and swift nest boxes with and without nest forms (case
346   study 3; *Newell, 2019*), they should be set to be equal or similar between treatment and control.
347   In contrast, when sample size is beyond control, e.g. it represents the number of sprouts in case
348   study 4 (*Andruk, 2014*), it is not possible to secure the equality of treatment and control sample
349   sizes. Some difference between treatment and control sample sizes is still acceptable, as in case
350   study 3, but its increase leads to inaccurate estimation of intervention effectiveness by the
351   magnitude of change. As a result, this study not only supports an appeal for balanced study

352   designs in ecological research (*Fletcher & Underwood, 2002*), but also indicates the conditions
353   when biases in estimation are possible.
354   The only situation when the relationships between treatment and control samples do not
355   matter is when the number of target outcomes in treatment sample is 0. This is the best outcome
356   for reduction-aimed interventions which is quantified by the relative risk and the magnitude of
357   change as the maximum reduction of target outcome by 100% (Table 1). Interventions applied to
358   reduce, *inter alia*, the rates of livestock depredation (*Eklund et al., 2017*; *van Eeden et al., 2018*),
359   deforestation (*Burivalova et al., 2019*) and poaching (*Kragt et al., 2020*) are all aimed at
360   reduction of target outcomes and their ultimate goal is to nullify them. However, such maximum
361   effectiveness cannot be assumed *a priori* even for interventions which often show excellent
362   results, like electric fences for reduction of livestock losses (*Khorozyan & Waltert, 2019a*;
363   *Khorozyan & Waltert, 2019b*), because the effectiveness is site-specific. Therefore, proper
364   setting of treatment and control sample sizes is essential anyway. These sample sizes, as well as
365   the numbers of target and alternative outcomes as their sub-samples (Fig. 1), should be explicitly
366   reported by researchers to allow independent evaluation of intervention effectiveness.
367   If designed treatment and control samples are unlikely to be similar, it is most reasonable to
368   expect the treatment sample to be smaller than the control sample due to significant efforts and
369   resources invested to the process of treatment (*Eklund et al., 2017*; *Snijders et al., 2019*). This
370   was confirmed by the first three case studies where sample sizes were determined by researchers
371   (*Knarrum et al., 2006*; *Sampaio, 2018*; *Newell, 2019*). In this situation, three scenarios are
372   possible: (1) if the number of target outcomes in treatment sample is lower than that in control
373   sample – overestimation by the magnitude of change; (2) if the number of target outcomes in
374   treatment sample is higher than that in control sample – underestimation by the magnitude of
375   change; and (3) if the numbers of target outcomes in treatment and control samples are equal –
376   the magnitude of change estimates no effectiveness and the relative risk indicates an increase in
377   target outcome (Table 1).
378   Why are overestimation and underestimation specified for the magnitude of change in
379   regard to the relative risk, but not vice versa? The point is that the magnitude of change does not
380   consider sample sizes and is thus a less reliable metric. The relative risk is a ratio of the
381   probability of target outcome with the intervention to that probability without the intervention
382   (*Eklund et al., 2017*; *Khorozyan & Waltert, 2019a*), so it represents a balanced metric accounting
383   for treatment and control sample sizes. It is one of the main effectiveness metrics in medicine
384   (*Stare & Maucort-Boulch, 2016*), a discipline which is well advanced in research of evidence-
385   based intervention effectiveness (*Pullin & Knight, 2001*; *Sutherland et al., 2004*; *Sutherland &
386   Wordley, 2017; Altringham, Berthinussen & Wordley, 2020*). In contrast, the magnitude of
387   change is based only on the difference between the numbers of target outcomes (*Jones &
388   Schmitz, 2009*; *Green et al., 2013*; *Miller et al., 2016*), yet sample sizes are critically important
389   and their ignorance can drastically change the effectiveness estimates. A vivid example is the
390   situation when the number of target outcomes in treatment sample is equal to that in control
391   sample. In this case, the magnitude of change fails to detect the effectiveness and quantifies it as

392 0, whereas the relative risk estimates the effectiveness to increase or decrease depending on
393 sample sizes (Table 1). In extreme cases of imbalanced samples, with high dissimilarity between
394 treatment and control samples, the magnitude of change can even show an opposite direction of
395 effectiveness in comparison with the relative risk. For example, imagine a fictitious case when
396 the numbers in Table 2 are changed as follows: number of killed ewes with bells – 1, number of
397 surviving ewes with bells – 7, number of killed ewes without bells – 3 and number of surviving
398 ewes without bells – 97. So, the treatment sample size is 8 and the control sample size is 100. In
399 this case, the magnitude of change shows a reduction of depredation rate by 66.7%, but the
400 relative risk shows an increase by 316.7%.
401   I hope that this study will stimulate more active implementation of well-designed scientific
402 and practical research in evidence-based conservation so that to make tangible and long-lasting
403 positive changes in the state of the environment.
404

Conclusions
405
406
407   Researchers should set equal treatment and control sample sizes in order to produce
408 identical estimates of intervention effectiveness by the relative risk and the magnitude of change.
409 When treatment and control samples are unequal, these estimates are biased if produced by the
410 magnitude of change and the relative risk should be used instead. As setting equal treatment and
411 control samples can be problematic for practical reasons, I encourage researchers and
412 practitioners to use the relative risk in estimation of intervention effectiveness. This will not take
413 additional efforts as both metrics are calculated from the same contingency table. Researchers
414 should report treatment and control sample sizes, and the numbers of target and alternative
415 outcomes within these samples, to let the readers evaluate intervention effectiveness
416 independently. Overall, this approach can strengthen study design and data analysis by scientists,
417 increase the reliability of results for practitioners, and facilitate decision-making by conservation
418 managers based on accurate information.
419

Acknowledgements
420
421
422   I thank A. Christie and an anonymous reviewer for providing insightful comments, which
423 significantly improved the quality of the manuscript.
424

References
425
426

427 Adams VM, Barnes M, Pressey RL. 2019. Shortfalls in conservation evidence: moving from
428   ecological effects of interventions to policy evaluation. *One Earth* 1:62-75 DOI:
429   10.1016/j.oneear.2019.08.017
430 Altringham JD, Berthinussen A, Wordley CFR. 2020. Generating, collating and using evidence
431   for conservation. In: Sutherland WJ, Brotherton PNM, Davies ZG, Ockendon N, Pettorelli

432   N, Vickery JA, eds. *Conservation research, policy and practice*. Cambridge: Cambridge
433   University Press, 48-62.
434 Andruk CM. 2014. Restoration of central Texas savanna and woodland: the effects of fire, deer
435   and invasive species on plant community trajectories. D. Phil. Thesis, University of Texas
436   at Austin.
437 Bruns A, Waltert M, Khorozyan I. 2020. The effectiveness of livestock protection measures
438   against wolves (*Canis lupus*) and implications for their co-existence with humans. *Global*
439   *Ecology and Conservation* 21:e00868 DOI: 10.1016/j.gecco.2019.e00868
440 Burivalova Z, Miteva D, Salafsky N, Butler RA, Wilcove DS. 2019. Evidence types and trends
441   in tropical forest conservation literature. *Trends in Ecology and Evolution* 34:669-679 DOI:
442   10.1016/j.tree.2019.03.002
443 Christie AP, Amano T, Martin PA, Shackelford GE, Simmons BI, Sutherland WJ. 2019. Simple
444   study designs in ecology produce inaccurate estimates of biodiversity responses. *Journal of*
445   *Applied Ecology* 56:2742-2754 DOI: 10.1111/1365-2664.13499
446 da Silva MX, Paviolo A, Tambosi LR, Pardini R. 2018. Effectiveness of protected areas for
447   biodiversity conservation: mammal occupancy patterns in the Iguaçu National Park, Brazil.
448   *Journal for Nature Conservation* 41:51-62 DOI: 10.1016/j.jnc.2017.11.001
449 Eklund A, López-Bao JV, Tourani M, Chapron G, Frank J. 2017. Limited evidence on the
450   effectiveness of interventions to reduce livestock predation by large carnivores. *Scientific*
451   *Reports* 7:2097 DOI: 10.1038/s41598-017-02323-w
452 Fletcher DJ, Underwood AJ. 2002. How to cope with negative estimates of components of
453   variance in ecological field studies. *Journal of Experimental Marine Biology and Ecology*
454   273:89-95 DOI: 10.1016/S0022-0981(02)00142-9
455 Green JMH, Larrosa C, Burgess ND, Balmford A, Johnston A. 2013. Deforestation in an African
456   biodiversity hotspot: extent, variation and the effectiveness of protected areas. *Biological*
457   *Conservation* 164:62-72 DOI: 10.1016/j.biocon.2013.04.016
458 Guthery FS, Bingham RL. 2007. A primer on interpreting regression models. *Journal of Wildlife*
459   *Management* 71:684-692 DOI: 10.2193/2006-285
460 Jones HP, Schmitz OJ. 2009. Rapid recovery of damaged ecosystems. *PLoS One* 4:e5653 DOI:
461   10.1371/journal.pone.0005653
462 Khorozyan I, Ghoddousi S, Soufi M, Soofi M, Waltert M. 2020. Studded leather collars are very
463   effective in protecting cattle from leopard (*Panthera pardus*) attacks. *Ecological Solutions*
464   *and Evidence* 1:e12013 DOI: 10.1002/2688-8319.12013
465 Khorozyan I, Waltert M. 2019a. A framework of most effective practices in protecting human
466   assets from predators. *Human Dimensions of Wildlife* 24:380-394 DOI:
467   10.1080/10871209.2019.1619883
468 Khorozyan I, Waltert M. 2019b. How long do anti-predator interventions remain effective?
469   Patterns, thresholds and uncertainty. *Royal Society Open Science* 6:190826 DOI:
470   10.1098/rsos.190826

471   Knarrum V, Sørensen OJ, Eggen T, Kvam T, Opseth O, Overskaug K, Eidsmo A. 2006. Brown
472   bear predation on domestic sheep in central Norway. *Ursus* 17:67-74 DOI: 10.2192/1537-
473   6176(2006)17[67:BBPODS]2.0.CO;2
474   Kragt ME, Hay E, Scheufele G, Bennett J, Renton M. 2020. Predicting the effectiveness of
475   community anti-poaching patrols for conserving threatened wildlife in the Lao PDR.
476   *Journal of Applied Ecology* 57:320-330 DOI: 10.1111/1365-2664.13527
477   McKillup S. 2012. *Statistics explained: an introductory guide for life scientists, 2nd ed.*
478   Cambridge: Cambridge University Press.
479   Miller JRB, Stoner KJ, Cejtin MR, Meyer TK, Middleton AD, Schmitz OJ. 2016. Effectiveness
480   of contemporary techniques for reducing livestock depredations by large carnivores.
481   *Wildlife Society Bulletin* 40:806-815 DOI: 10.1002/wsb.720
482   Nakagawa S, Cuthill IC. 2007. Effect size, confidence interval and statistical significance: a
483   practical guide for biologists. *Biological Reviews* 82:591-605 DOI: 10.1111/j.1469-
484   185X.2007.00027.x
485   Newell D. 2019. A test of the use of artificial nest forms in common swift *Apus apus* nest boxes
486   in southern England. *Conservation Evidence* 16: 24-26.
487   Pullin AS, Cheng SH, Cooke SJ, Haddaway NR, Macura B, McKinnon MC, Taylor JJ. 2020.
488   Informing conservation decisions through evidence synthesis and communication. In:
489   Sutherland WJ, Brotherton PNM, Davies ZG, Ockendon N, Pettorelli N, Vickery JA, eds.
490   *Conservation research, policy and practice*. Cambridge: Cambridge University Press, 114-
491   128.
492   Pullin AS, Knight TM. 2001. Effectiveness in conservation practice: pointers from medicine and
493   public health. *Conservation Biology* 15:50-54 DOI: 10.1111/j.1523-1739.2001.99499.x
494   Pullin AS, Knight TM, Stone DA, Charman K. 2004. Do conservation managers use scientific
495   evidence to support their decision-making? *Biological Conservation* 119:245-252 DOI:
496   10.1016/j.biocon.2003.11.007
497   Sampaio M. 2018. Green turtles on the island of Cavalos (Guinea-Bissau). Abundance, nest
498   success and experimental nest protection. M. Sc. Thesis, University of Lisbon.
499   Schmidt BR, Arlettaz R, Schaub M, Lüscher B, Kröpfli M. 2019. Benefits and limits of
500   comparative effectiveness studies in evidence-based conservation. *Biological Conservation*
501   236:115-123 DOI: 10.1016/j.biocon.2019.05.033
502   Smith RK, Dicks LV, Mitchell R, Sutherland WJ. 2014. Comparative effectiveness research: the
503   missing link in conservation. *Conservation Evidence* 11:2-6.
504   Snijders L, Greggor AL, Hilderink F, Doran C. 2019. Effectiveness of animal conditioning
505   interventions in reducing human-wildlife conflict: a systematic map protocol.
506   *Environmental Evidence* 8:10 DOI: 10.1186/s13750-019-0153-7
507   Stare J, Maucort-Boulch D. 2016. Odds ratio, hazard ratio and relative risk. Metodološki zvezki
508   13:59-67.

509   Sutherland WJ, Pullin AS, Dolman PM, Knight TM. 2004. The need for evidence-based
510   conservation. *Trends in Ecology and Evolution* 19:305-308 DOI:
511   10.1016/j.tree.2004.03.018
512   Sutherland WJ, Taylor NG, MacFarlane D, Amano T, Christie AP, Dicks LV, Lemasson AJ,
513   Littlewood NA, Martin PA, Ockendon N, Petrovan SO, Robertson RJ, Rocha R,
514   Shackelford GE, Smith RK, Tyler EHM, Wordley CFR. 2019. Building a tool to overcome
515   barriers in research-implementation spaces: the conservation evidence database. *Biological*
516   *Conservation* 238:108199 DOI: 10.1016/j.biocon.2019.108199
517   Sutherland WJ, Wordley CFR. 2017. Evidence complacency hampers conservation. *Nature*
518   *Ecology and Evolution* 1:1215-1216 DOI: 10.1038/s41559-017-0244-1
519   Toomey AH, Knight AT, Barlow J. 2017. Navigating the space between research and
520   implementation in conservation. *Conservation Letters* 10:619-625 DOI:
521   10.1111/conl.12315
522   Treves A, Krofel M, Ohrens O, van Eeden LM. 2019. Predator control needs a standard of
523   unbiased randomized experiments with cross-over design. *Frontiers in Ecology and*
524   *Evolution* 7:462 DOI: 10.3389/fevo.2019.00462
525   van Eeden LM, Crowther MS, Dickman CR, Macdonald DW, Ripple WJ, Ritchie EG, Newsome
526   TM. 2017. Managing conflict between large carnivores and livestock. *Conservation*
527   *Biology* 32:26-34 DOI: 10.1111/cobi.12959
528   van Eeden LM, Eklund A, Miller JRB, López-Bao JV, Chapron G, Cejtin MR, Crowther MS,
529   Dickman CR, Frank J, Krofel M, Macdonald DW, McManus J, Meyer TK, Middleton AD,
530   Newsome TM, Ripple WJ, Ritchie EG, Schmitz OJ, Stoner KJ, Tourani M, Treves A.
531   2018. Carnivore conservation needs evidence-based livestock protection. *PLoS Biology*
532   16:e2005577 DOI: 10.1371/journal.pbio.2005577
533   Wan H, Chizinski CJ, Dolph CL, Vondracek B, Wilson BN. 2010. The impact of rare taxa on a
534   fish index of biotic integrity. *Ecological Indicators* 10:781-788 DOI:
535   10.1016/j.ecolind.2009.12.006
536   Westgate MJ, Haddaway NR, Cheng SH, McIntosh EJ, Marshall C, Lindenmayer DB. 2018.
537   Software support for environmental evidence synthesis. *Nature Ecology and Evolution*
538   2:588-590 DOI: 10.1038/s41559-018-0502-x
539   Wielgus RB, Peebles KA. 2014. Effects of wolf mortality on livestock depredation. *PLoS One* 9:
540   e113505 DOI: 10.1371/journal.pone.0113505
541   Woodroffe R, Frank LG, Lindsey PA, ole Ranah SMK, Romañach S. 2007. Livestock husbandry
542   as a tool for carnivore conservation in Africa's community rangelands: a case-control
543   study. *Biodiversity and Conservation* 16:1245-1260 DOI: 10.1007/s10531-006-9124-8

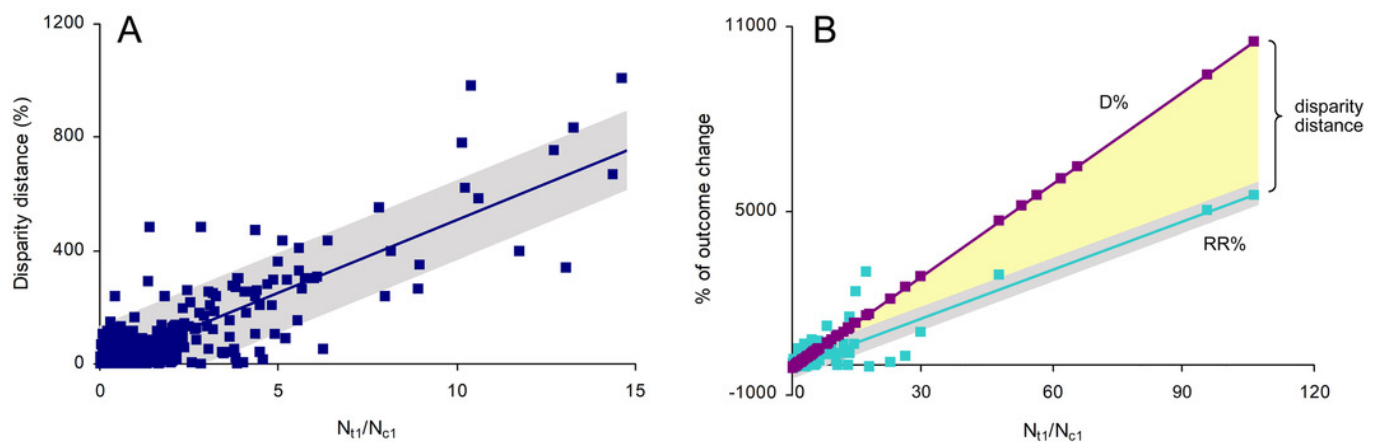
Figure 1

The 2×2 contingency table used for estimation of intervention effectiveness.

		Outcomes		
		No. target outcomes	No. alternative outcomes	
Intervention	Treatment = Intervention present	N_{t1}	N_{t2}	$N_{t1} + N_{t2} = N_t$, treatment sample size
	Control = Intervention absent	N_{c1}	N_{c2}	$N_{c1} + N_{c2} = N_c$, control sample size
		$N_{t1} + N_{c1} = N_1$, total number of target outcomes	$N_{t2} + N_{c2} = N_2$, total number of alternative outcomes	

Figure 2

The effects of the ratio of the number of target outcomes in treatment sample to that in control sample (N_{t1}/N_{c1}) on (A) disparity distance between the relative risk RR% and the magnitude of change D% and on (B) RR% and D% separately.

The line on figure (A) and the RR% line on figure (B) are indicated with their 95% confidence intervals.

The scheme of disparities between the magnitude of change (D%) and the relative risk (RR%) in estimating the effectiveness of conservation interventions depending on relationships between the numbers of target outcomes and sample sizes.

See the contingency table in Fig. 1 for more details.

1   Table 1. The scheme of disparities between the magnitude of change (D%) and the relative risk
2   (RR%) in estimating the effectiveness of conservation interventions depending on relationships
3   between the numbers of target outcomes and sample sizes. See the contingency table in Fig. 1 for
4   more details.

Numbers of target outcomes[1]	Treatment and control sample sizes[2]	Relationships between D% and RR%[3]	Explanation
$N_{t1} < N_{c1}$	$N_t > N_c$	$\|D\%\| < \|RR\%\|$	Underestimation by D%
	$N_t < N_c$	$\|D\%\| > \|RR\%\|$	Overestimation by D%
	$N_t = N_c$	$D\% = RR\%$	Same estimate by both metrics
$N_{t1} > N_{c1}$	$N_t > N_c$	$\|D\%\| > \|RR\%\|$	Overestimation by D%
	$N_t < N_c$	$\|D\%\| < \|RR\%\|$	Underestimation by D%
	$N_t = N_c$	$D\% = RR\%$	Same estimate by both metrics
$N_{t1} = N_{c1}$	$N_t > N_c$	$D\% = 0, D\% > RR\%$	No effect by D%, decrease of target outcome by RR%
	$N_t < N_c$	$D\% = 0, D\% < RR\%$	No effect by D%, increase of target income by RR%
	$N_t = N_c$	$D\% = RR\% = 0$	No effect of intervention
$N_{t1} = 0$	-	$D\% = RR\% = -100\%$	Maximum reduction of target outcome by 100%

7   [1] Abbreviations: N_{t1} – number of target outcomes in treatment sample; N_{c1} – number of target
8   outcomes in control sample.
9   [2] Abbreviations: N_t – treatment sample size and N_c – control sample size.
10 [3] The absolute estimates are provided because D% and RR% can be both positive and negative.

The numbers of ewes with and without bell collars killed by brown bears (*Ursus arctos*) in Norway in 1994 (*Knarrum et al., 2006*).

1   Table 2. The numbers of ewes with and without bell collars killed by brown bears (*Ursus arctos*)
2   in Norway in 1994 (*Knarrum et al., 2006*).

	No. killed	No. not killed	Total
With bells	10	9	19
Without bells	7	27	34
Total	17	36	53

The numbers of green turtle (*Chelonia mydas*) nests with and without metal nets predated by Nile monitors (*Varanus niloticus*) in Guinea-Bissau in 2016 (*Sampaio, 2018*).

1 Table 3. The numbers of green turtle (*Chelonia mydas*) nests with and without metal nets
2 predated by Nile monitors (*Varanus niloticus*) in Guinea-Bissau in 2016 (*Sampaio, 2018*).

	No. predated	No. not predated	Total
With metal nets	3	23	26
Without metal nets	19	45	64
Total	22	68	90

The numbers of nest boxes with and without nest forms occupied by swifts (*Apus apus*) in the UK in 2009-2018 (*Newell, 2019*).

1   Table 4. The numbers of nest boxes with and without nest forms occupied by swifts (*Apus apus*)
2   in the UK in 2009-2018 (*Newell, 2019*).

	No. occupied	No. not occupied	Total
With nest forms	30	32	62
Without nest forms	12	68	80
Total	42	100	142

The numbers of hardwood sprouts in plots with and without prescribed fire in the USA in 2009-2012 (*Andruk, 2014*).

1   Table 5. The numbers of hardwood sprouts in plots with and without prescribed fire in the USA
2   in 2009-2012 (*Andruk, 2014*).

	No. sprouts post-treatment	No. sprouts pre-treatment	Total
With fire	669	247	916
Without fire	152	92	244
Total	821	339	1160