Methods in Ecology and Evolution
British Ecological Society

Methods in Ecology and Evolution 2015, **6,** 576–583 doi: 10.1111/2041-210X.12338

Determining hormone metabolite concentrations when enzyme immunoassay accuracy varies over time

Eve Davidian†, Sarah Benhaiem*†, Alexandre Courtiol, Heribert Hofer, Oliver P. Höner and Martin Dehnhard

Leibniz Institute for Zoo and Wildlife Research, Alfred-Kowalke-Strasse 17, D-10315 Berlin, Germany

Summary

1. Enzyme immunoassays (EIAs) are widely used to quantify concentrations of hormone metabolites. Modifications in laboratory conditions may affect the accuracy of metabolite concentration measurements and lead to misinterpretations when results of different accuracy are combined for a statistical analysis. This issue is of great relevance to studies in behavioural and evolutionary ecology because these usually aim at understanding how hormone concentrations vary between individuals, environments or experimental conditions.

2. We present a method based on re-assaying a subset of samples to standardize hormone metabolite concentrations when changes in EIA accuracy occur. We used glucocorticoid metabolite concentrations (fGMCs) measured in faeces of spotted hyaenas (*Crocuta crocuta*) between 2011 and 2013 with a previously validated EIA. Changes in accuracy were assessed by monitoring the metabolite concentration of faecal control 'pools' that were systematically assayed with faecal samples. A cluster analysis on these pools identified two distinct sample sets with different EIA accuracy; 'Cluster 1' and 'Cluster 2'. We then re-assayed all samples of Cluster 1 ($n = 138$) with an EIA accuracy similar to that of Cluster 2 and fitted a linear regression to the remeasured fGMCs against the initial fGMCs to predict fGMCs in Cluster 2. To determine the minimum number of samples to re-assay that allows reliable predictions, we assessed the variation in the quality of model predictions by fitting linear regressions on decreasing numbers of re-assayed samples. This revealed that re-assaying 27 samples would be sufficient to generate reliable predictions considering our data set.

3. To test the robustness of our method, we fitted a new linear regression to 27 randomly chosen samples and used its equation to standardize all fGMCs of Cluster 1. The standardized fGMCs were similar to the remeasured fGMCs, and the regression on 27 samples was as effective at standardizing fGMCs as the regression fitted on the complete data set.

4. Our standardization method permits the combination of results of different accuracy. It is a simple and reliable alternative to the costly, time-consuming and often impractical re-assaying of complete sample sets that can be applied to a wide variety of species and sample types.

Key-words: accuracy, bias, binding reaction, control pool, enzyme immunoassay, long-term study, spotted hyaena, standardization, steroid metabolite

Introduction

Methods to measure concentrations of steroid hormone metabolites in urine and faeces have become an essential part of studies in evolutionary ecology and conservation. They have been applied to many taxa to investigate key topics such as the interplay between steroid hormones and social or sexual behaviour (Rasmussen *et al.* 2008; Benhaiem *et al.* 2013) and the physiological response of endangered species to disturbance (Rolland *et al.* 2012). Because the collection of urine and faeces does not involve potentially stressful procedures such as the manipulation or immobilisation of study animals, these methods are particularly useful to monitor adrenocortical activity over time and for studies on free-ranging animals (Hofer & East 1998; Touma & Palme 2005; Landys, Goymann & Slagsvold 2011; Rolland *et al.* 2012; Benhaiem *et al.* 2013).

Hormone metabolite concentrations are most commonly quantified using enzyme immunoassays (EIAs) (Touma & Palme 2005). In indirect, competitive EIAs, the metabolites in faecal or urine samples compete with a known amount of tracer (e.g. a steroid hormone conjugated with a peroxidase enzyme) for the binding sites of a hormone-specific antibody. The proportion of bound tracer generates a 'response', which is read photometrically and expressed as optical density. The metabolite concentration in a sample is then quantified by relating the optical density to a calibrated dose–response curve generated by standards of known hormone concentration (Wild 2013). Because the chemical structure of metabolites and their binding affinity towards the antibody usually differs from that of their native hormone contained in standards and tracer, there often is a bias between measured and 'true' metabolite concentration; this bias defines the 'accuracy' of an EIA

*Correspondence author. E-mail: benhaiem@izw-berlin.de
†Contributed equally to this work

(Wild 2013). Such a bias is negligible if it remains constant for all assayed samples and if, as in most studies in behavioural and evolutionary ecology, the main interest lies in relative differences in concentrations between individuals, environments or experimental conditions (Lynch *et al.* 2003; Brown, Walker & Steinman 2004). If, however, EIA accuracy substantially changes during the course of the study, measured metabolite concentrations in samples are not directly comparable and combining them in statistical models would lead to misinterpretations and erroneous conclusions.

Changes in EIA accuracy may occur for various reasons. EIAs involve binding reactions that are sensitive to laboratory conditions such as room temperature and exposure to light during incubation (responsible for 'edge' or 'well-to-well' effects; Watson *et al.* 2013). Other potential sources of variation in binding reactions are modifications in protocols such as changes in the concentration of antibody and tracer, replacement of the antibody, standards or other reagents when they are used up, expire or when commercial kits are discontinued, changes of equipment, and switches in laboratory personnel (Shekarchi *et al.* 1984; Jones *et al.* 1995; Noble *et al.* 2008; Wasser *et al.* 2010; Watson *et al.* 2013). Because the native hormone in standards and tracer and metabolites in samples differ in their chemical properties, their binding reactions may be affected differently by variation in laboratory protocol and conditions, thereby inducing changes in the accuracy of metabolite concentrations (Watson *et al.* 2013).

To track changes in EIA accuracy and other characteristics of EIA performance, control parameters such as the responses in blank wells, the standard concentrations associated with relative binding sensitivities (i.e. concentrations at 10, 20, 50, 80 and 90% of binding) and metabolite concentrations in urine or faecal control solutions or 'pools' are routinely monitored (Brown, Walker & Steinman 2004; Wild 2013). Pools are commonly used to assess the intra-assay and interassay repeatability or 'precision' of measurements of metabolite concentrations. Substantial changes in the concentration of pools additionally indicate changes in the relationship between measurements of metabolites and standards, that is, changes in EIA accuracy (Gill, Hayes & Sluss 2003). One possibility to avoid non-comparable results due to changes in accuracy is to (re-)assay all samples together and within a short period of time. Re-assaying large data sets for each new research question, however, is costly in time, manpower, sample material and money and may not always be feasible, for example when samples are depleted.

Here, we present a method to standardize results when changes in EIA accuracy occur, based on the re-assaying of a subset of samples. We establish this method using glucocorticoid metabolite concentrations measured in faeces (fGMCs) of spotted hyaenas (*Crocuta crocuta*) collected in the Ngorongoro Crater, Tanzania, as part of a long-term research project (Höner *et al.* 2007, 2010). We demonstrate that our method effectively standardizes metabolite concentrations and allows comparison of measurements obtained when EIA accuracy varies.

Methods and Results

COLLECTION AND TREATMENT OF FAECAL SAMPLES

We collected 483 faecal samples from 272 free-ranging spotted hyaenas between 2002 and 2013. Faeces were collected immediately after defaecation, mixed, subsampled and stored in liquid nitrogen until transported to Germany on dry ice where they were stored at −80°C until further processing. Faecal subsamples were freeze-dried (for 49–70 h) with a Lyovac-GT2 lyophilisator (Hürth, Germany). Aliquots of 0·1 g were extracted with 0·9 mL of 90% methanol for 30 min, centrifuged, and the supernatant (typically 0·7 mL) diluted 1:1 with distilled water. Faecal samples and extracts were stored at −80°C between treatments.

ASSAY OF FAECAL SAMPLES

Faecal extracts were assayed in three batches by two technicians; one technician assayed extracts in July 2011 (n = 71 extracts, 5 plates) and September 2011 (n = 67 extracts, 5 plates) and the other in July 2013 (n = 345 extracts, 13 plates). We quantified fGMCs using an 'in-house' cortisol-3-CMO competitive EIA that was validated for spotted hyaenas and demonstrated a high affinity of the antibody with cortisol metabolites, the ability of the antibody to measure natural fluctuation in metabolite concentrations and a high precision of measurement (Benhaiem *et al.* 2012). We used microtitre plates coated with a polyclonal antibody raised in rabbits against cortisol-3-CMO-BSA and cortisol-3-CMO-peroxidase as tracer (for more details on EIA protocol, see Benhaiem *et al.* 2012). Calibrated standard curves were prepared by serial 1:2 dilutions of a cortisol stock solution and ranged from 0·2 to 100 pg 20 μL^{-1}. Calibration curves were fitted using Akima's spline interpolation (Akima 1970). The approximately linear range of the calibration curve (i.e. the section between 20 and 80% of binding of the tracer or 'binding sensitivity') was used to estimate fGMCs in samples using Magellan software (version 2.6; Tecan Group Ltd., Männedorf, Switzerland). Faecal samples with concentrations exceeding this range (typically >25 pg 20 μL^{-1}) were diluted to provide precise quantification of metabolite concentrations. Final fGMCs were obtained by multiplying the measured raw fGMCs by their corresponding factor of dilution and expressed as ng g^{-1} of dry faecal matter.

We used two faecal control pools with relatively high and low metabolite concentrations (hereafter: 'high pool' and 'low pool') to monitor intra-assay and interassay precision and potential changes in EIA accuracy. Stock solutions of pools and standards were renewed several times (but never simultaneously) during the course of the study. Coefficients of variation (CV) between old and new stocks never exceeded 10%, thus complying with the commonly accepted interassay coefficient of variation ($CV_{interassay}$) of 20% (for details on this criterion, see the section on EIA performance and for the CV formula, see Appendix S1 in Supporting Information), and confirming that renewals of stocks were not associated with changes in pool concentrations nor with shifts in EIA binding sensitivity. Assay plates were subdivided following a design that was constant throughout the study, with specific wells assigned to standard solutions, pools, blank controls, and faecal extracts, respectively. All extracts and controls were assayed in duplicate and, as typically recommended (Wild 2013), measurements were only accepted when duplicated values did not differ by more than 5% from their mean (i.e. $CV_{intra\text{-}assay} \leq 5\%$). The concentrations of antibody and tracer were changed during the course of the study, but all other parameters of the experimental procedure and equipment were maintained constant.

DATA ANALYSIS

Statistical procedures were performed using R software version 3.1.0 (R Development Core Team 2013). Results are quoted as mean ± standard deviation (SD), probabilities are for two-tailed tests, the threshold for significance was set at 5%, and 95% percentile confidence intervals ($CI_{95\%}$) were calculated using a bootstrap method with 100 000 iterations (R package 'boot'; Canty & Ripley 2014). For ordinary least squares (OLS) linear regressions, including analyses of covariance (ANCOVA), the distribution of residuals did not significantly deviate from normality (Shapiro–Wilk tests) and the variances were homoscedastic (Breusch-Pagan tests and residuals plots; R package 'car'; Fox & Weisberg 2011).

EIA PERFORMANCE DURING THE STUDY PERIOD

To assess whether the samples assayed during the entire study had comparable fGMCs, we calculated the $CV_{interassay}$ of the fGMCs of the pools across all 23 plates. The $CV_{interassay}$ of pools (mean = 58·4 ± 20·2%) exceeded the commonly applied criterion of 20% (e.g. Goymann *et al.* 1999; Bales *et al.* 2005; Ganswindt *et al.* 2005; Behie, Pavelka & Chapman 2010), indicating substantial variation in the accuracy of metabolite measurements during the study.

We then assessed separately for each of the three batches EIA precision, stability of the accuracy of measurements, analytical sensitivity, quantitative resolution and binding sensitivity. $CV_{intra\text{-}assay}$ and $CV_{interassay}$ of pools did not exceed 5% and 20%, respectively, indicating that the precision of the EIA was high and the accuracy remained stable within each batch of measurements (see Appendix S1 in Supporting Information). The results also indicated that the EIA maintained a high analytical sensitivity and quantitative resolution throughout the study. The binding sensitivity of the EIA was similar in July 2011 (range of standard concentrations at 10–90% of binding: 0·5–28·9 pg 20 μL^{-1}) and September 2011 (range: 0·6–30·5 pg 20 μL^{-1}) but lower in July 2013 (range: 0·9–40·5 pg 20 μL^{-1}; Fig. 1).

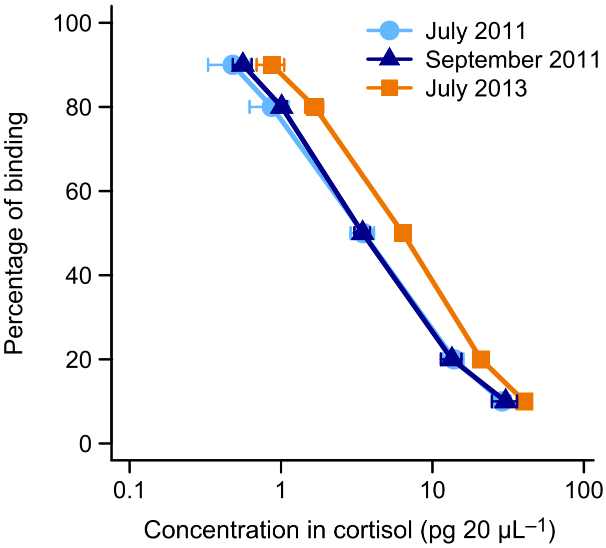

Fig. 1. Relationship between the percentage of binding of the tracer and measured cortisol concentration in standards. Symbols correspond to the mean ± SD concentration in cortisol at 10, 20, 50, 80 and 90% of binding of the tracer, for July 2011 (*n* = 5 standard curves), September 2011 (*n* = 5 standard curves) and July 2013 (*n* = 13 standard curves).

We also tested for interference with non-antigenic material in samples because such 'matrix effects' can disrupt the relationship between measurements of metabolites and hormone standards when faecal extracts are diluted. We applied two tests of parallelism which compared the slope of the calibration curve with that of the displacement curve obtained from serial dilutions of faecal extracts (Kemeny & Challacombe 1988). One test was performed in July 2011 using two faecal samples ('A' and 'B') that were extracted in 2011 and a second test was performed in July 2013 using two faecal samples ('C' and 'D') that were extracted in 2011 and 2012. Parallelism was validated for all four faecal extracts (ANCOVA, *P*-value for comparison of the slopes: *P* = 0·08 for extract A; *P* = 0·86 for B; *P* = 0·15 for C and *P* = 0·51 for D), demonstrating that there were no matrix effects on our measurements and that the bias between measurements of metabolites and standards was constant throughout the range of dilution of faecal extracts.

ESTABLISHMENT OF THE STANDARDIZATION PROCEDURE

The standardization procedure consists of (i) identifying 'clusters' of samples assayed with similar EIA accuracy and assigning the reference cluster, (ii) choosing and re-assaying a subset of samples with the EIA accuracy of the reference cluster, (iii) modelling the relationship between initial and remeasured metabolite concentrations, (iv) testing the predictive performance of the model and (v) standardizing the metabolite concentrations of all samples from the cluster. The following sections detail how we established the method using our data set on fGMCs in spotted hyenas.

Identifying clusters of samples assayed with similar EIA accuracy and assigning the reference cluster

To identify plates that contain pools of similar concentrations (i.e. similar accuracy) and determine which and how many faecal samples may need to be re-assayed and standardized, we conducted a cluster analysis on all measurements of fGMCs of the high and low pools simultaneously. We performed a hierarchical clustering using Ward's agglomeration method on the dissimilarity matrix of Euclidean distances between the fGMCs of pools from the 23 plates (R core package 'stats'; Ward 1963; Murtagh & Legendre 2014). This analysis identified two distinct clusters, referred to as Cluster 1 and Cluster 2 (Fig. 2). Cluster 1 comprised the 10 plates (*n* = 138 samples) assayed in July 2011 and September 2011 and Cluster 2 the 13 plates (*n* = 345 samples) assayed in July 2013. An alternative analysis conducted on the fGMCs of each pool separately revealed similar results. To verify that each cluster conformed to the generally accepted interassay variation in precision and accuracy of $CV_{interassay}$ ≤20%, we calculated the $CV_{interassay}$ of the pools in Cluster 1 and Cluster 2. The $CV_{interassay}$ of pools in the two clusters each conformed to the level of acceptance, indicating a stable EIA accuracy within each cluster (Fig. 2). We assigned Cluster 2 as the reference cluster for the re-assaying of samples and standardization of fGMCs because at the time of re-assaying EIA accuracy corresponded to that of Cluster 2 (see following section). The fGMCs of low and high pools increased by a factor of 7·5 and 2·7, respectively, between Cluster 1 and Cluster 2.

Re-assaying samples with the EIA accuracy of the reference cluster

To establish our method, we re-assayed all 138 faecal samples of Cluster 1 within a few days after assaying the 345 samples of Cluster 2, using the same solutions of standards and pools, and applying the same EIA

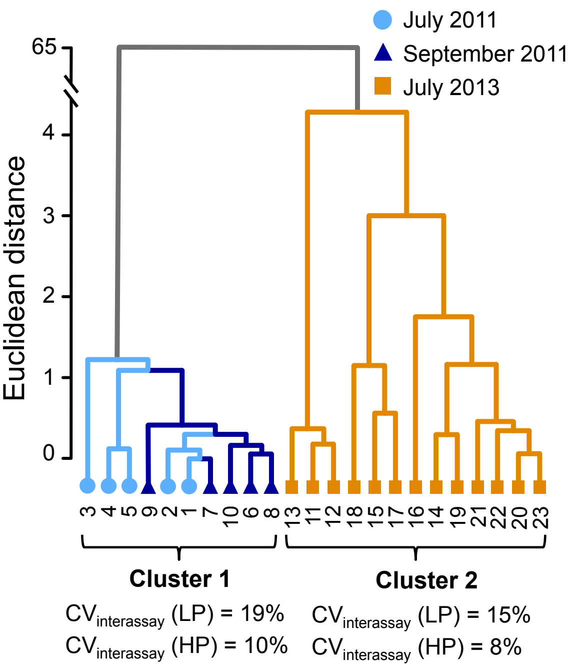

Fig. 2. Dendrogram showing the hierarchical clustering of the concentration of faecal pools assayed on 23 plates. The two clusters of faecal pools identified by the analysis are referred to as Cluster 1 and Cluster 2. $CV_{interassay}$ (LP) and $CV_{interassay}$ (HP) correspond to the interassay coefficient of variation for the low and high pool, respectively. Numbers (from 1 to 23) below the dendrogram refer to the code of the plate.

protocol as for Cluster 2. Optical densities of standards and concentrations of pools run with the re-assayed samples were similar to those of standards and pools run in Cluster 2 ($CV_{interassay}$ <20%), confirming a stable EIA accuracy and binding sensitivity between Cluster 2 and re-assaying. To avoid errors associated with different extraction procedures and dilutions of sample extracts, all re-assays were performed using the same sample extracts and dilutions as in Cluster 1.

Modelling the relationship between initial and remeasured concentrations using the complete data set

We modelled the relationship between the fGMCs initially measured in Cluster 1 ($fGMC_{initial}$, as x) and the fGMCs remeasured with the accuracy of Cluster 2 ($fGMC_{remeasured}$, as y) using an OLS linear regression on raw measurements, that is, before multiplying fGMCs by their corresponding dilution factor (for a comparison with an alternative linear model, see Appendix S1 in Supporting Information). The resulting equation was as follows:

$$fGMC_{remeasured} = 4{\cdot}22 + 1{\cdot}33 \times fGMC_{initial} \qquad \text{eqn 1}$$

This model accounted well for the variation in remeasured fGMCs (adjusted $r^2 = 0{\cdot}72$; $n = 138$).

Cross-validating the model using the complete data set

We assessed the predictive performance of the model using a cross-validation procedure (R package 'DAAG'; Maindonald & Braun 2014) that divided the data set into three subsets of equal size. Alternately, two subsets were grouped and used as 'training' sets to fit an OLS linear regression while the remaining subset was used as a 'test' set to assess the reliability of predictions on an independent subset of samples.

Following our tolerated variation in precision and accuracy of repeated measurements of $CV_{interassay}$ ≤20%, we considered model predictions to be reliable if the difference between predicted fGMCs and their matched remeasured fGMCs did not exceed 20% (i.e. CV_{fit} ≤20%). We further considered a model to have a satisfactory predictive performance when at least 70% of samples had a CV_{fit} ≤20%. The cross-validation showed that 86·2% of samples (120 out of 138 samples) conformed to our criterion of CV_{fit}.

Estimating the minimum number of samples to re-assay

To estimate the minimum number of samples required to obtain reliable predictions, we generated data sets of decreasing size (i.e. from 138 to 10 samples) by choosing randomly, without replacement, a given number of faecal samples among the complete data set of 138 samples. We fitted and cross-validated OLS linear regressions on these data sets. The random sampling, model fitting and cross-validation procedures were reiterated 10 000 times for each data set. Results of the cross-validation of these models are illustrated in Fig. 3 for two different criteria of prediction reliability (i.e. CV_{fit} ≤20% and CV_{fit} ≤10%; see also Table S2 in Appendix S1 in Supporting Information). The smallest data set to reach our threshold for model acceptance (i.e. 70% of CV_{fit} ≤20%) was 27 samples (Fig. 3).

Applying and validating the standardization procedure with a subset of 27 samples

To test the effectiveness of the standardization based on a subset of samples, we randomly chose 27 samples from our complete data set of 138 samples and fitted an OLS linear regression to their initial and remeasured fGMCs (adjusted $r^2 = 0{\cdot}65$; $n = 27$; Fig. 4). The resulting equation was as follows:

$$fGMC_{remeasured} = 4{\cdot}88 + 1{\cdot}21 \times fGMC_{initial} \qquad \text{eqn 2}$$

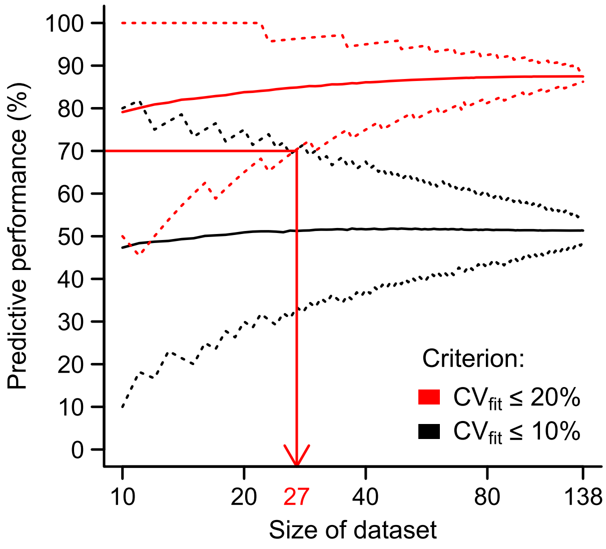

Fig. 3. Variation in the predictive performance of OLS linear regressions with decreasing size of the data set. Results indicate the mean (solid lines) and 95% confidence interval (dotted lines) percentage of faecal samples with a coefficient of variation (CV_{fit}) within the boundaries of CV_{fit} ≤20% and CV_{fit} ≤10%, calculated with 10 000 simulations. The red line with arrow indicates the smallest number of samples to re-assay (here, $n = 27$) to obtain at least 70% of samples with a CV_{fit} ≤20%. The x-axis is displayed on a logarithmic scale.

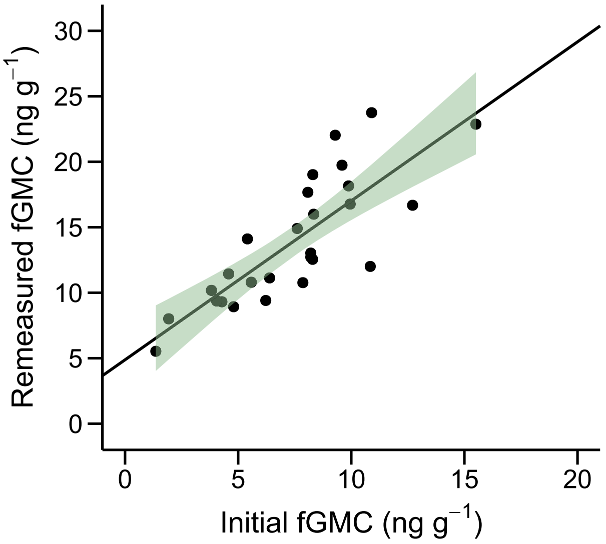

Fig. 4. Relationship between initial and remeasured faecal glucocorticoid metabolite concentrations (fGMC) for a subset of 27 faecal samples. The black line is the OLS linear regression fitted to predict fGMCs in Cluster 2 (eqn 2: $\text{fGMC}_{\text{remeasured}} = 4{\cdot}88 + 1{\cdot}21 \times \text{fGMC}_{\text{initial}}$, adjusted $r^2 = 0{\cdot}65$). The area shaded in green represents the 95% confidence interval of the fit.

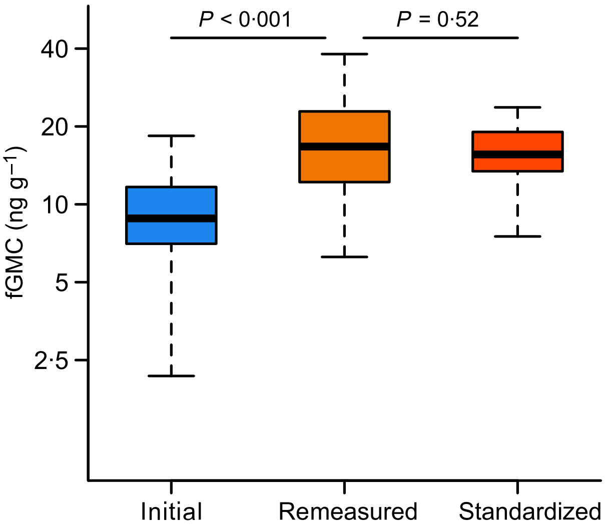

Fig. 5. Faecal glucocorticoid metabolite concentrations (fGMC) initially measured in Cluster 1, remeasured with the accuracy of Cluster 2 and standardized on Cluster 2 (n = 138). Boxes encompass interquartile ranges (first to third quartiles around the median), horizontal lines inside boxes represent medians, and whiskers are at 1·5 times the interquartile ranges. The y-axis is displayed on a logarithmic scale.

The cross-validation of the model indicated that 88·9% of samples (24 out of 27 samples) had a $\text{CV}_{\text{fit}} \leq 20\%$, confirming that our subset was large enough to perform reliable standardization. Moreover, the Pearson product–moment correlation coefficient between predicted and remeasured fGMCs for the model based on the subset of samples was high (eqn 2, $r = 0{\cdot}81$, $n = 27$) and did not differ from the correlation obtained for the model based on the complete data set (eqn 1, $r = 0{\cdot}85$, $n = 138$; Fisher's r to Z transformation; $Z = -0{\cdot}58$, $P = 0{\cdot}56$).

We standardized the fGMCs of all 138 samples of Cluster 1 using eqn 2 and rescaled the raw standardized concentrations into final concentrations by multiplying them by their dilution factor. To test the effect of the variation in EIA accuracy on the quantification of fGMCs, assess the potential risk of misinterpreting results when fGMCs of different accuracy are combined, and verify that our standardization procedure effectively reduced such risk, we compared the fGMCs of the 138 samples that were remeasured with the accuracy of Cluster 2 to (i) their matched fGMCs initially measured in Cluster 1 and (ii) their matched fGMCs after standardization on Cluster 2. As expected by the observed increase in pool concentrations, the fGMCs remeasured with the accuracy of Cluster 2 (median = 16·7 ng g^{-1}) were significantly higher than their matched fGMCs measured in Cluster 1 (median = 8·8 ng g^{-1}; Wilcoxon's signed-rank test, $V = 9591$, $P < 0{\cdot}0001$; median of between-group differences = 7·44 ng g^{-1}, $\text{CI}_{95\%}$ = 6·76–8·33 ng g^{-1}; Fig. 5), but did not significantly differ from their matched standardized fGMCs (median = 15·6 ng g^{-1}; $V = 4494$, $P = 0{\cdot}52$; median of between-group differences = −0·30 ng g^{-1}, $\text{CI}_{95\%}$ = −0·84 to 0·42 ng g^{-1}; Fig. 5).

Discussion

Our results confirm that changes in EIA accuracy can bias measurements of metabolite concentrations. Measurements of varying accuracy should therefore be standardized before being combined for statistical analysis. We showed that our method reduced the differences in fGMCs caused by the change in EIA accuracy between two clusters to a non-significant value, rendering all measurements of the study comparable with each other. We further demonstrate the reliability of the standardization procedure when only a small subset of samples is re-assayed. To our knowledge, this is the first method to standardize metabolite concentrations when changes in accuracy occur within a given EIA.

The method involves simple statistical procedures, applies relevant and widely accepted criteria in endocrinology and can be generalized to cases when two or more clusters require standardization (see Box 1 for a summary of the procedure). Appropriate consideration should be given to the number of samples to re-assay and the regression method applied to model the relationship between initial and remeasured metabolite concentrations. The minimum number of samples to re-assay is study specific and depends on various factors such as the maximum intra-assay and interassay variation in precision and accuracy that is tolerated (here, 5% and 20%, respectively), the threshold for model acceptance and the dispersion of sample metabolite concentrations (Linnet 1999; Brown, Walker & Steinman 2004). Here, we proposed a threshold for model acceptance of 70% of samples complying with our criterion of prediction reliability, and this indicated a minimum subset of 27 samples to be re-assayed. The model fitted to 27 samples was as effective at standardizing fGMCs as the model for the complete data set, confirming that this threshold was sufficient in our case. A different threshold may be better suited to other data sets and scientific questions. Note, however, that even when fitted to the complete data set, the mean predictive performance of our model never exceeded 87·5% (see Table

Box 1. Procedure to standardize sample metabolite concentrations measured by EIAs

Step 1: Identify clusters of samples assayed with similar EIA accuracy

- Define the EIA accuracy at the time of standardization as the reference accuracy. If unknown, run a plate with standards and pools to determine it.
- Conduct a hierarchical cluster analysis on the concentration of pools of all plates and assign clusters based on the resulting dendrogram.
- Calculate the interassay coefficient of variation (CV) of pools for each cluster; subdivide clusters with CVs exceeding the criterion for similar EIA accuracy (here, 20%). Repeat until all clusters have CVs that satisfy the criterion.
- Define the cluster of samples assayed with the reference accuracy as the reference cluster and standardize samples from all other clusters on this cluster.

Note: The following steps describe the procedure to standardize one cluster. If Step 1 indicates that more clusters should be standardized, repeat Step 2 to Step 5 for each cluster.

Step 2: Choose and re-assay a subset of samples

- Choose a subset of samples that is representative of all samples of the cluster and that can be re-assayed at the same dilution as when initially assayed.
- Re-assay the subset within a few days after the cluster analysis to ensure that the EIA accuracy of the subset and reference cluster is, similar.

Step 3: Model the relationship between initial and remeasured metabolite concentrations

- Model the relationship between initial (x) and remeasured (y) concentrations of the subset of samples using raw measurements, that is, before multiplying them by their dilution factor and retrieve the resulting equation (i.e. intercept and slope).

Step 4: Test the predictive performance of the model

- Cross-validate the model and retrieve the predicted metabolite concentrations of the samples in the subset.
- Set a criterion for prediction reliability that corresponds to the criterion for similar EIA accuracy (here, 20%) and compute the CVs of the predicted and remeasured metabolite concentrations for each sample in the subset.
- Set a threshold for satisfying model predictive performance (here, 70%) and calculate the percentage of samples that conform to the criterion for prediction reliability. If this threshold is not reached, re-assay more samples (i.e. restart from Step 2).

Step 5: Standardize the metabolite concentrations

- Standardize the concentrations of all other samples of the cluster using the equation obtained in Step 3.
- Rescale the standardized raw concentrations into final concentrations by multiplying them by their dilution factor.

Note: R programing codes for each step of the standardization procedure are provided as Supporting Information in Appendix S2 (Supporting Information).

S2, Appendix S1 in Supporting Information). We suggest starting with a subset of approximately 27 samples and cross-validating the fitted model to assess *a posteriori* whether predictions are reliable or whether additional samples need to be re-assayed. The approximate number of additional samples required to reach the chosen threshold can be estimated with the help of Table S2 (Appendix S1 in Supporting Information) for two different criteria of prediction reliability.

Whether the samples in the subset should be chosen in a random manner from the complete data set or randomly within some stratification may depend on the scientific question and the distribution of metabolite concentrations in the cluster. If measurements reflect different treatments or categories of individuals (e.g. diet, age, sex, social status), samples should be randomly chosen within each treatment or category of individuals (Pocock & Simon 1975). To avoid introducing errors owing to different dilution factors, we further recommend fitting a model on measured raw metabolite concentrations, that is, not corrected for their dilution. The chosen subset should therefore be restricted to samples that can be re-assayed at the same dilution as their initial measurement. The factor of change in concentration of the high and low pools between two clusters can be used to estimate the highest and lowest initial raw metabolite concentration in samples that is likely to fall within the linear range of the calibration curve if remeasured at the same dilution.

The relationship between initial and remeasured concentrations may in most cases be best described by a linear regression, but other methods (e.g. polynomial) may give a better fit depending on how concentrations changed along with the change in EIA accuracy. Alternatively, nonparametric regression techniques such as splines (Green & Silverman 1993) may be applied. In our study, we only accepted measurements when their duplicated values differed by <5%. Applying a simple OLS linear regression was as effective at standardizing metabolite concentrations as a more complex model that explicitly incorporated measurement errors (see Appendix S1 in Supporting Information). If a larger discrepancy between duplicated measurements is tolerated, OLS linear models may have a lower predictive power than models that incorporate measurement errors on both axes.

The standardization relies on the ability to track changes in EIA accuracy using the metabolite concentration of pools. It is thus important to be able to dismiss the effect of erroneous preparation of hormone standard and pool solutions on metabolite measurements. To facilitate this, we highly recommend preparing new stock solutions of pools and standards before the old stocks are depleted and assaying them together on a transition plate to ensure that new solutions give similar results to the old ones (Brown, Walker & Steinman 2004). When old and new stock solutions of pools are prepared based on different faecal or urinary samples, for example when the original samples are depleted, such a procedure allows to adjust the dilution of new solutions to match the concentration of old pools or, alternatively, to calculate a factor of change between old and new pools. The evaluation of the performance of our EIA indicated that the EIA remained highly precise and sensitive throughout the study. The validation of parallelism at the beginning and end of the study confirmed the absence of matrix effects that could have been associated with the extraction procedure, denaturation of the antibody or changes in the structure of metabolites in faecal extracts over time. Renewal of stock solutions of standards and faecal pools did not coincide with changes in the optical densities of standards or with changes in the metabolite concentration of pools. Finally, the switch in laboratory personnel that occurred between the two clusters is unlikely to be the cause of the change in EIA accuracy because several technicians used the same EIA during our study and all experienced a similar change in binding sensitivity and accuracy. The observed change in accuracy and binding sensitivity between the two clusters thus most probably resulted from adjustments in the concentration of antibody and tracer and potential (uncontrolled) fluctuations in environmental conditions (e.g. room temperature).

Our assessment of the standardization procedure was based on re-assaying a subset of samples after some time had elapsed (18 months and more). Ageing of faecal samples, that is, the latencies between sample collection, storage, extraction and assaying, may alter metabolite concentration owing to naturally occurring faecal bacteria that may decompose steroid metabolites after defaecation (Möstl & Palme 2002). Applying appropriate treatment, storage and extraction procedures can stabilise hormone metabolites for long periods of time, possibly for many years. Here, we applied the recommended treatment and storage procedures for faecal steroids (Khan *et al.* 2002; Terio *et al.* 2002; Hunt & Wasser 2003; Lynch *et al.* 2003; Millspaugh & Washburn 2004; Kalbitzer & Heistermann 2013) and found that the increase in fGMC of sample extracts was consistent with that of pools, suggesting that the age of extracts had no or only a minor influence on the measurements.

Quantifying the absolute accuracy of metabolite measurements is difficult because the chemical structure and binding affinity of metabolites are usually unknown. We therefore cannot determine whether the initial or the remeasured metabolite concentrations are more accurate. This is usually of little relevance in studies in ecology and evolution where a similar level of EIA accuracy and comparable measurements are more important than a high EIA accuracy in absolute terms (Lynch *et al.* 2003). Methods have been developed within the context of clinical studies and studies in conservation medicine to compare and harmonise measurements of plasma hormones obtained with different EIAs (Bidlingmaier & Freda 2010). However, the global application of such harmonisation methods may be limited because they are often based on the systematic re-assaying of complete sets of samples, are not aimed at predicting standardized concentrations and rarely consider intra-EIA changes in accuracy, sensitivity or precision that would affect the harmonisation procedure over time (Müller *et al.* 2011).

Our standardization method may be particularly useful for collaborative projects that share the laboratory workload between different facilities and are likely to experience variation in EIA performance and accuracy, and for long-term and longitudinal studies that typically deal with large data sets and may not be able to re-assay all samples whenever new samples are collected or a new research question is investigated. Moreover, because this method only requires the re-assaying of a subset of samples, it allows the standardization of the initial measurements of samples that are no longer available. This can significantly increase sample sizes, enhance the power of statistical analyses and allow the inclusion of a larger number of covariates in statistical models, which may be important for a better understanding of complex processes.

Ethical statement

All procedures were performed in accordance with the requirements of the Leibniz Institute for Zoo and Wildlife Research Ethics Committee on Animal Welfare.

Acknowledgements

We are grateful to Marion L. East, Susanne Pribbenow, Wolfgang Goymann and three anonymous reviewers for constructive comments, Stephanie Kramer-Schadt and Thomas Petzoldt for help with *R* code programming, and Katrin Paschmionka, Sylvia Schulz van Endert, Heidrun Barleben, Marlies Rohleder and Dagmar Thierer for their valuable technical assistance in the laboratory. We thank the Tanzania Commission for Science and Technology, the Tanzania Wildlife Research Institute and the Ngorongoro Conservation Area Authority for permission to conduct the study. This study was financed by the Deutsche Forschungsgemeinschaft (grant HO 2498/4-1) and the Leibniz Institute for Zoo and Wildlife Research.

Data accessibility

The data are archived in figshare: http://dx.doi.org/10.6084/m9.figshare.1298131.

References

Akima, H. (1970) A new method of interpolation and smooth curve fitting based on local procedures. *Journal of the Association for Computing Machinery*, **17**, 589–602.

Bales, K.L., French, J.A., Hostetler, C.M. & Dietz, J.M. (2005) Social and reproductive factors affecting cortisol levels in wild female golden lion tamarins (*Leontopithecus rosalia*). *American Journal of Primatology*, **67**, 25–35.

Behie, A.M., Pavelka, M.S. & Chapman, C.A. (2010) Sources of variation in fecal cortisol levels in howler monkeys in Belize. *American Journal of Primatology*, **72**, 600–606.

Benhaiem, S., Dehnhard, M., Bonanni, R., Hofer, H., Goymann, W., Eulenberger, K. & East, M.L. (2012) Validation of an enzyme immunoassay for the measurement of faecal glucocorticoid metabolites in spotted hyenas *Crocuta crocuta*. *General and Comparative Endocrinology*, **178**, 265–271.

Benhaiem, S., Hofer, H., Dehnhard, M., Helms, J. & East, M.L. (2013) Sibling competition and hunger increase allostatic load in spotted hyaenas. *Biology Letters*, **9**, 20130040.

Bidlingmaier, M. & Freda, P.U. (2010) Measurement of human growth hormone by immunoassays: current status, unsolved problems and clinical consequences. *Growth Hormone & IGF Research*, **1**, 19–25.

Brown, J.L., Walker, S. & Steinman, K. (2004) *Endocrine Manual for Reproductive Assessment of Domestic and Non-domestic Species*. Conservation and Research Center, Smithsonian's National Zoological Park, Front Royal, Virginia.

Canty, A. & Ripley, B. (2014) *boot: Bootstrap R (S-Plus) Functions*. R package version 1.3-13.

Fox, J. & Weisberg, S. (2011) *An {R} Companion to Applied Regression*, 2nd edn. R package version 2.0-22. Sage, Thousand Oaks CA.

Ganswindt, A., Rasmussen, H.B., Heistermann, M. & Hodges, J.K. (2005) The sexually active states of free-ranging male African elephants (*Loxodonta africana*): defining musth and non-musth using endocrinology, physical signals, and behavior. *Hormones and Behavior*, **47**, 83–91.

Gill, S., Hayes, F.J. & Sluss, P.M. (2003) Issues in endocrine immunoassay. *Contemporary Endocrinology: Handbook of Diagnostic Endocrinology* (eds J.E. Hall & L.K. Nieman), pp. 1–22. Humana Press Inc., Totowa, NJ.

Goymann, W., Möstl, E., Van't Hof, T., East, M.L. & Hofer, H. (1999) Noninvasive fecal monitoring of glucocorticoids in spotted hyenas, *Crocuta crocuta*. *General and Comparative Endocrinology*, **114**, 340–348.

Green, P.J. & Silverman, B.W. (1993) *Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach*. Chapman & Hall/CRC Press, London.

Hofer, H. & East, M.L. (1998) Biological conservation and stress. *Advances in the Study of Behavior*, **27**, 405–525.

Höner, O.P., Wachter, B., East, M.L., Streich, W.J., Wilhelm, K., Burke, T. & Hofer, H. (2007) Female mate-choice drives the evolution of male-biased dispersal in a social mammal. *Nature*, **448**, 798–802.

Höner, O.P., Wachter, B., Hofer, H., Wilhelm, K., Thierer, D., Trillmich, F., Burke, T. & East, M.L. (2010) The fitness of dispersing spotted hyaena sons is influenced by maternal social status. *Nature Communications*, **1**, 60.

Hunt, K.E. & Wasser, S.K. (2003) Effect of long-term preservation methods on fecal glucocorticoid concentrations of grizzly bear and African elephant. *Physiological and Biomedical Zoology*, **76**, 918–928.

Jones, G., Wortberg, M., Kreissig, S.B., Hammock, B.D. & Rocke, D.M. (1995) Sources of experimental variation in calibration curves for enzyme-linked immunosorbent assay. *Analytica Chemica Acta*, **313**, 197–207.

Kalbitzer, U. & Heistermann, M. (2013) Long-term storage effects in steroid metabolite extracts from baboon (*Papio* sp.) faeces – a comparison of three commonly applied storage methods. *Methods in Ecology and Evolution*, **4**, 493–500.

Kemeny, D.M. & Challacombe, S.J. (eds) (1988) *ELISA and Other Solid Phase Immunoassays: Theoretical and Practical Aspects*. John Wiley & Sons Ltd., New York.

Khan, M.Z., Altmann, J., Isani, S.S. & Yu, J. (2002) A matter of time: evaluating the storage of fecal samples for steroid analysis. *General and Comparative Endocrinology*, **128**, 57–64.

Landys, M.M., Goymann, W. & Slagsvold, T. (2011) Rearing conditions have long-term consequences for stress responsiveness in free-living great tits. *General and Comparative Endocrinology*, **174**, 219–224.

Linnet, K. (1999) Necessary sample size for method comparison studies based on regression analysis. *Clinical Chemistry*, **45**, 882–894.

Lynch, J.W., Khan, M.Z., Altmann, J., Njahira, M.N. & Rubenstein, N. (2003) Concentrations of four fecal steroids in wild baboons: short-term storage conditions and consequences for data interpretation. *General and Comparative Endocrinology*, **132**, 264–271.

Maindonald, J.H. & Braun, W.J. (2014) DAAG: Data Analysis and Graphics Data and Functions. R package version 1.20.

Millspaugh, J.J. & Washburn, B.E. (2004) Use of fecal glucocorticoid metabolite measures in conservation biology research: considerations for application and interpretation. *General and Comparative Endocrinology*, **138**, 189–199.

Möstl, E. & Palme, R. (2002) Hormones as indicators of stress. *Domestic Animal Endocrinology*, **23**, 67–74.

Müller, A., Scholz, M., Blankenstein, O., Binder, G., Pfaffle, R., Korner, A. *et al.* (2011) Harmonization of growth hormone measurements with different immunoassays by data adjustment. *Clinical Chemistry and Laboratory Medicine*, **49**, 1135–1142.

Murtagh, F. & Legendre, P. (2014) Ward's hierarchical agglomerative clustering method: which algorithms implement Ward's criterion? *Journal of Classification*, **31**, 274–295.

Noble, J.E., Wang, L., Cerasoli, E., Knight, A.E., Porter, R.A., Gray, E. *et al.* (2008) An international comparability study to determine the sources of uncertainty associated with a non-competitive sandwich fluorescent ELISA. *Clinical Chemistry and Laboratory Medicine*, **46**, 1033–1045.

Pocock, S.J. & Simon, R. (1975) Sequential treatment assignment with balancing for prognostic factors in the controlled clinical trial. *Biometrics*, **31**, 103–115.

R Development Core Team (2013) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna.

Rasmussen, H.B., Ganswindt, A., Douglas-Hamilton, I. & Vollrath, F. (2008) Endocrine and behavioral changes in male African elephants: linking hormone changes to sexual state and reproductive tactics. *Hormones and Behavior*, **54**, 539–548.

Rolland, R.M., Parks, S.E., Hunt, K.E., Castellote, M., Corkeron, P.J., Nowacek, D.P., Wasser, S.K. & Kraus, S.D. (2012) Evidence that ship noise increases stress in right whales. *Proceedings of the Royal Society B*, **279**, 2363–2368.

Shekarchi, I.C., Sever, J.L., Lee, Y.J., Castellano, G. & Madden, D.L. (1984) Evaluation of various plastic microtiter plates with measles, toxoplasma, and gamma globulin antigens in enzyme-linked immunosorbent assays. *Journal of Clinical Microbiology*, **19**, 89–96.

Terio, K.A., Brown, J.L., Moreland, R. & Munson, L. (2002) Comparison of different drying and storage methods on quantifiable concentrations of fecal steroids in the cheetah. *Zoo Biology*, **21**, 215–222.

Touma, C. & Palme, R. (2005) Measuring fecal glucocorticoid metabolites in mammals and birds: the importance of validation. *Annals of the New York Academy of Sciences*, **1046**, 54–74.

Ward, J.H. Jr (1963) Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, **58**, 236–244.

Wasser, S.K., Azkarate, J.C., Booth, R.K., Hayward, L., Hunt, K., Ayres, K. *et al.* (2010) Non-invasive measurements of thyroid hormone in feces of a diverse array of avian and mammalian species. *General and Comparative Endocrinology*, **168**, 1–7.

Watson, R., Munro, C., Edwards, K.L., Norton, V., Brown, J.L. & Walker, S.L. (2013) Development of a versatile enzyme immunoassay for non-invasive assessment of glucocorticoid metabolites in a diversity of taxonomic species. *General and Comparative Endocrinology*, **186**, 16–24.

Wild, D. (2013) *The Immunoassay Handbook*, 4th edn. Elsevier, Oxford.

Received 20 November 2014; accepted 13 January 2015
Handling Editor: Diana Fisher

Supporting Information

Additional Supporting Information may be found in the online version of this article.

Appendix S1. Enzyme immunoassay performance, alternative regression method and variation in model predictive performance with decreasing size of data set.

Appendix S2. R programming codes to apply the standardization procedure.