PLOS | ONE

RESEARCH ARTICLE

Bayesian Estimation of Small Effects in Exercise and Sports Science

Kerrie L. Mengersen[1,2]*, Christopher C. Drovandi[1,2], Christian P. Robert[3], David B. Pyne[4,5], Christopher J. Gore[4,5,6]

1 Science and Engineering Faculty, Mathematical Sciences, and Institute for Future Environments, Queensland University of Technology, Brisbane, Australia, **2** Australian Research Council Centre of Excellence in Mathematical and Statistical Frontiers in Big Data, Big Models and New Insights, Brisbane, Australia, **3** Ceremade, Universite Paris Dauphine, Paris, France, **4** Australian Institute of Sport, Canberra, Australia, **5** Research Institute for Sport and Exercise, University of Canberra, Bruce, ACT, Australia, **6** Exercise Physiology Laboratory, Flinders University of South Australia, Bedford Park, South Australia

* k.mengersen@qut.edu.au

Citation: Mengersen KL, Drovandi CC, Robert CP, Pyne DB, Gore CJ (2016) Bayesian Estimation of Small Effects in Exercise and Sports Science. PLoS ONE 11(4): e0147311. doi:10.1371/journal.pone.0147311

Editor: Cathy W.S. Chen, Feng Chia University, TAIWAN

Received: August 25, 2015

Accepted: December 31, 2015

Published: April 13, 2016

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Funding: The authors have no support or funding to report.

Competing Interests: The authors have declared that no competing interests exist.

Abstract

The aim of this paper is to provide a Bayesian formulation of the so-called magnitude-based inference approach to quantifying and interpreting effects, and in a case study example provide accurate probabilistic statements that correspond to the intended magnitude-based inferences. The model is described in the context of a published small-scale athlete study which employed a magnitude-based inference approach to compare the effect of two altitude training regimens (live high-train low (LHTL), and intermittent hypoxic exposure (IHE)) on running performance and blood measurements of elite triathletes. The posterior distributions, and corresponding point and interval estimates, for the parameters and associated effects and comparisons of interest, were estimated using Markov chain Monte Carlo simulations. The Bayesian analysis was shown to provide more direct probabilistic comparisons of treatments and able to identify small effects of interest. The approach avoided asymptotic assumptions and overcame issues such as multiple testing. Bayesian analysis of unscaled effects showed a probability of 0.96 that LHTL yields a substantially greater increase in hemoglobin mass than IHE, a 0.93 probability of a substantially greater improvement in running economy and a greater than 0.96 probability that both IHE and LHTL yield a substantially greater improvement in maximum blood lactate concentration compared to a Placebo. The conclusions are consistent with those obtained using a 'magnitude-based inference' approach that has been promoted in the field. The paper demonstrates that a fully Bayesian analysis is a simple and effective way of analysing small effects, providing a rich set of results that are straightforward to interpret in terms of probabilistic statements.

Introduction

A key interest in sports science is the estimation and evaluation of small effects, such as the difference in finishing times between world-class athletes, or the impact of exercise training and/or lifestyle interventions such as dietary changes or sleep behaviors on performance [1]. While such an interest is not confined to this context [2], there are some features of sports science that make accurate and relevant estimation of small effects particularly challenging. Two such challenges are small sample sizes when dealing with international-standard, elite-level athletes and frequent small true between-individual differences in competitive performance. The issue of dealing with small sample sizes in studies has drawn comment in the fields of both medicine [3, 4] and sports science [5].

These issues have been addressed by a number of sports science researchers. For example, Batterham and Hopkins (2006) challenged the traditional method of making an inference based on a p-value derived from a hypothesis test, arguing that it is confusing, potentially misleading and unnecessarily restrictive in its inferential capability [6]. The authors suggested alternative is to focus on the confidence interval as a measure of the uncertainty of the estimated effect, and examine the proportion of this interval that overlaps pre-defined magnitudes that are clinically or mechanistically relevant. As illustration, Batterham and Hopkins identify 'substantially positive', 'trivial' and 'substantially negative' magnitudes, as well as more finely graded magnitudes. The authors then translate these proportions to a set of likelihood statements about the magnitude of the true effect.

Batterham and Hopkins justify their suggested approach and corresponding inferences by drawing an analogy between their method and a Bayesian construction of the problem. In particular, they claim that their approach is approximately Bayesian based on no prior assumption about the distribution of the true parameter values. This has drawn criticism by a number of authors, such as Barker and Schofield (2008) who–rightly–point out that the approach is *not* Bayesian, and that the assumed priors in an analogous Bayesian approach may indeed be informative [7]. More recently, Welsh and Knight (2014) further criticised the approach of Batterham and Hopkins and suggested that relevant statistical approaches should use either confidence intervals or a fully Bayesian analysis [8].

The aim of this paper is to provide a Bayesian formulation of the method proposed by Batterham and Hopkins (2006) and provide a range of probabilistic statements that parallel their intended magnitude-based inferences. The models described here can be expanded as needed to address other issues. For further exposition, the model is described in the context of a small-scale athlete study authored by Humberstone-Gough and co-workers [9], which employed Batterham and Hopkins' approach to compare the effect of two altitude training regimens (live high train low, and intermittent hypoxic exposure) on running performance and blood measurements of elite triathletes.

Methods

General model

Both Bayesian and frequentist approaches require specification of a statistical model for the observed data, which contains a number of parameters that need to be estimated. Bayesian methods are different from frequentist approaches in that the parameters are treated as random variables. That is, they are considered as having true, but unknown, values and are thus described by a (posterior) probability distribution that reflects the uncertainty associated with how well they are known, based on the data. The posterior distribution is obtained by multiplying the likelihood, which describes the probability of observing the data given specified values

of the parameters, and the prior distribution(s), which encapsulates beliefs about the probability of obtaining those parameter values independently of the data. These priors may be developed using a range of information sources including previous experiments, historical data and/or expert opinion. Alternatively, they may be so-called uninformative or vague distributions, to allow inferences to be driven by the observed data.

This study describes a simple statistical model that might be considered in the context of examining small effects in sports science and also some possible prior distributions that might be placed on the parameters of this model. Some extensions to the model are considered in a later section.

Suppose that there are G treatment groups. For the gth group ($g = 1,\ldots,G$), let n_g denote the total number of individuals in the group, $y_{i(g)}$ denote an observed effect of interest for the ith individual in the group ($i = 1,\ldots,n_g$), y_g denote the set of observations in the group, $\bar{y}_g$ and s_g^2 denote respectively the sample mean and sample standard deviation of all the observed responses from the group, and $\nu_g = n_g - 1$ denote the degrees of freedom. For example, in the following case study, there are $G = 3$ groups (training regimens); $y_{i(g)}$ is the difference between the post- and pre-treatment measurements for a selected response for the ith athlete in the gth training regimen, and $\bar{y}_g$ is the average difference for that group.

Assume that an observation $y_{i(g)}$ is Normally distributed around a group mean μ_g, with a group-specific variance σ_g^2, i.e.:

$$y_{i(g)} \sim \text{Normal}(\mu_g, \sigma_g^2) \tag{1}$$

A vague prior density is adopted for the pair of parameters (μ_g, σ_g^2) [10] so that:

$$p(\mu_g, \sigma_g^2) \propto \sigma_g^{-2} \tag{2}$$

(where $\propto$ denotes proportional to). Based on [1] and [2], the posterior conditional distributions for μ_g and σ_g^2 are given by

$$\mu_g|\sigma_g^2, y_g \sim \text{N}(\bar{y}_g, \frac{\sigma_g^2}{n_g}) \tag{3}$$

$$\sigma_g^2|y_g \sim \text{Inverse}\chi^2(\upsilon_g, s_g^2). \tag{4}$$

The marginal posterior distribution for μ_g can be shown to have a t distribution on ν_g degrees of freedom: [10]

$$\mu_g|y_g \sim t_g(\bar{y}_g, \frac{s_g^2}{n_g}) \tag{5}$$

so that

$$(\mu_g - \bar{y}_g)/(s_g^2/\sqrt{n_g})|y_g \sim t_{\upsilon g} \tag{6}$$

Relationship with frequentist results

The marginal posterior distributions for σ_g^2 and μ_g, based on the data, are given by Eqs (4) and (5), respectively. Because of the choice of the vague prior (Eq (2)), these distributions can be shown to be closely related to analogous distributions for the (appropriately scaled) sufficient

statistics, given μ_g and σ_g^2, based on frequentist sampling theory: [10]

$$\upsilon_g s_g^2 | \sigma_g^2 \sim \chi_{\upsilon g}^2 \tag{7}$$

$$(\bar{y}_g - \mu_g)/(s_g^2/\sqrt{n_g}) \sim t_{\upsilon g}. \tag{8}$$

Estimation of values of interest

A range of posterior estimates (conditional on the available data) arising from the model may be of interest, including:

1. the mean and standard deviation for each group (e.g., each training regimen in the study), given by μ_g and σ_g^2, respectively
2. the difference between the group means: $\delta_{kl} = \mu_k - \mu_l$ and the associated standard deviation of this difference, σ_{kl}
3. a $(1-\alpha)$% credible interval (CI) for a measure of interest, θ, say, such that there is a posterior probability $(1-\alpha)$ that θ lies in this interval (e.g., θ could be the mean of group 2, i.e., $\theta = \mu_2$, and a 95% CI of (3.1, 4.2), for instance, indicates that the probability that μ_2 is between 3.1 and 4.2, given the data, is 0.95), which is a much more direct statement than is possible under a frequentist approach
4. Cohen's d [11] for the difference between two groups, given by $d_{kl} = \delta_{kl}/\sigma_{kl}$ when comparing groups k and l, $k \neq l$
5. the probability that Cohen's d exceeds a specified threshold such as a 'smallest worthwhile change' (*SWC*,[6]), given by $\Pr(d_{kl} > \text{SWC})$ or $\Pr(d_{kl} < -\text{SWC})$, depending on whether d_{kl} is positive or negative, respectively
6. the predicted outcome of each individual under each training regimen, regardless of whether or not they have participated in that training, obtained from Eq (1), with an estimate of the corresponding uncertainty of this prediction
7. the ranks of each individual under each training regimen, with corresponding uncertainty in these orderings.

Given the data y_g for each group (and hence the sufficient statistics $\bar{y}_g$ and s_g^2), it is straightforward to use Eqs (4) and (5) to compute posterior estimates μ_g and σ_g^2, and other probabilities of interest. An alternative, simple approach is to simulate values of interest using Eqs (3) and (4) iteratively, employing a form of Markov chain Monte Carlo (MCMC) [12]. A more technical explanation of this approach including the Gibbs sampling techniques is provided by Geman and Geman [13]. At each iteration, a value of σ_g^2 is simulated from Eq (4) and then a value of μ_g given that value of σ_g^2 is simulated from Eq (3). This process is repeated a large number of times. The simulated values can be used to compute other measures (e.g. $\exp(\mu_1 - \mu_2)$ if this is of interest), indicators $\text{I}(\mu_1 > c)$ or $\text{I}(\mu_1 > \mu_2)$ and so on. Then $\text{E}(\exp(\mu_1 - \mu_2))$, $\Pr(\mu_1 > c)$ and $\Pr(\mu_1 > \mu_2)$ can be estimated (where E denotes expectation) as the respective averages of these values over all of the iterations. Similarly, at each iteration, the simulated parameter values can be input into Eq (1) to obtain predicted values of y for each individual under each regimen, and the individuals can be ranked with respect to their predicted outcome. The posterior distributions for individual predicted outcomes, and the probability distribution for the ranks, are computed from the respective values obtained from the set of iterations.

The Cohen's *d* is a standardized effect size estimate, calculated as the difference between two means divided by the corresponding standard deviation. While there are many effect size estimators, Cohen's *d* is one of the most common since it is appropriate for comparing between the means of distinctly different group and it has appealing statistical properties; for example it has a well known distribution and is maximum likelihood estimator [14].

Model extensions

The model described above can be easily extended in a range of ways. Three such extensions are considered here. The first extension is that other prior distributions can be considered instead of Eq (2) above. For example, another common approach is to assign a normal distribution for the group means,

$$\mu_g \sim \text{Normal}(\text{M}, \text{V}) \tag{9}$$

and a Uniform distribution for the standard deviations,

$$\sigma_g \sim \text{Uniform}(0, \text{R}), \tag{10}$$

where M and V denote the mean and variance of the normal distribution, respectively, and R is the upper bound of the uniform distribution. Alternatives to the uniform are the half-normal or half-Cauchy. If the sample sizes within groups are small and little is known *a priori* about the comparative variability of measurements within and between the groups, then σ_g^2 can be imprecisely estimated; to avoid this, the individual variances be replaced by a common variance, σ^2 say.

There are many ways of setting the values of M, V and R. For example, if there is no prior information about these values and if the groups are considered to be independent, this can be reflected by specifying very large values of V and R, relative to the data. This means that the priors in Eqs (9) and (10) will have negligible weight in the posterior estimates of the group means μ_g and variances σ_g^2. If V is sufficiently large, the value of M will not matter, so it is commonly set to 0 in this case. Alternatively, the groups can be perceived as having their own characteristics (described by μ_g and σ_g^2) but also being part of a larger population with an overall mean M and variance V. This random effects model is very common as it helps to accommodate outliers and improve estimation of small groups. Another alternative is to use other information to set the values of M, V and R. This information can include results of previous similar experiments, published estimates, expert opinion, and so on. Depending on the problem and the available information, different values of M, V and R can be defined for the different groups. The Bayesian framework can be very helpful in providing a mechanism for combining these sources of information in a formal manner.

The second extension is that the model described in Eq (1) can be expanded to include explanatory variables that can help to improve the explanation or prediction of the response. This is the model that is adopted in the case study described below, where the explanatory variables comprise the group label and a covariate reflecting training-induced changes. For this purpose, Eq (1) is extended as follows:

$$y_i = x_i'\beta + \varepsilon \tag{11}$$

where the explanatory variables and their regression coefficients are denoted by x and β, respectively, and ε_i describes the residual between the observation y_i and its predicted value $x_i'\beta$. Note that the superscript $'$ denotes the transpose. It is common to assume that $\in_i \sim$Normal

$(0,\sigma^2)$. Normally distributed priors are placed on the parameters in this regression model:

$$\beta \sim \text{Normal}_k(b_0, \text{B}_0^{-1}); \sigma^2 \sim \text{Gamma}(c_0/2,\ \ d_0/2) \quad (12)$$

where k represents the number of parameters, Normal_k indicates a k-dimensional Gaussian distribution and Gamma indicates a Gamma distribution described by shape and scale parameters, in this case given by constants c_0 and d_0.

An uninformative prior specification can be defined for β by setting zero values for the mean vector b_0 and precision matrix B_0. Similarly, negligible prior information about the magnitude of the residuals is reflected by setting small values for c_0 and d_0 in the distribution for σ^2 [15].

An alternative, popular formulation is to use Zellner's g-prior, whereby the variance of the prior for β is defined in terms of the variance for the data. More explicitly, b is specified as a multivariate normal distribution with a covariance matrix that is proportional to the inverse Fisher information matrix for β, given by $g(x^Tx)^{-1}$. This is an elegant way of specifying the 'information' contained in the prior, relative to that contained in the data: the value of g is analogous to the 'equivalent number of observations' that is contributed to the analysis by the prior [16, 17].

The third extension is the choice of the response y. This depends on the aim of the analysis, biological and other contextual knowledge of the problem, and the available data. The residuals are assumed to have a normal distribution with a mean of zero, and normally distributed priors can be defined as the difference between an individual's post-training and pre-training measurements, the difference of the logarithms of these measurements, the relative difference between the pairs of measurements (i.e. (post-pre)/pre) or some other context-relevant transformation.

Case Study

The Bayesian approach described above was applied to a study by Humberstone-Gough *et al.* [9] who used a two-period (pre-post) repeated measures design to compare the effects of three training regimens 'Live High Train Low' altitude training (LHTL), 'Intermittent Hypoxic Exposure' (IHE) and 'Placebo' on running performance and blood characteristics. The study comprised eight subjects (elite male triathletes) in each regimen, and had one dropout in the LHTL group. Although ten running and blood variables were considered in the original study; three variables with the most complete data are selected here for illustration: hemoglobin mass (Hbmass, units of grams), submaximal running economy (RunEcon, units of L O_2.min^{-1}) and maximum blood lactate concentration (La-max, units of mmol/L). The authors also employed a covariate reflecting training-induced changes, namely the percent change in weekly training load from pre- to during-camp for each individual athlete. The data used for the analyses are shown in S1 Table (data extracted from original study of Humberstone-Gough et al (2013 and provided by co-author Gore).

Casting this study in terms of the models described above, there are $G = 3$ groups denoting the training regimens (Placebo by $g = 1$; IHE by $g = 2$; LHTL by $g = 3$). Letting pre_i and post_i denote respectively the pre- and post-treatment measurements for the ith individual, an (unscaled) effect of interest, y_i, was defined in terms of the difference between the pairs of measurements:

$$y_i = \text{post}_i - \text{pre}_i. \quad (13)$$

A log transformation was adopted in the original analysis by Humberstone-Gough *et al.* [9] but was not performed in the analysis described below, as there was insufficient information in the observed data to strongly motivate a transformation of the measurements, particularly after adjusting for the covariate reported by Humberstone-Gough *et al.* (comparative summary plots not shown). However, it is acknowledged that this decision was based purely on the

available data and there may be compelling biological or experimental reasons for choosing the log (or other) scale; for example, under this transformation covariates can be considered to have multiplicative rather than additive effects on the original response. On the one hand, retaining the original scale allows for more direct interpretation of the results. On the other, if the underlying assumptions are not met, the inferences based on the results must be treated with caution. In this study, the premise was adopted of not transforming unless there is a compelling domain-specific or statistical reason to do so. Hence the decision was made not to take a log transformation of the data as other authors have suggested–a statistical decision–and to consider a relative change in performance as well as an absolute difference–a domain-based decision since this measure is of interest to sports scientists. A similar issue arises about the inclusion of covariates in a small sample analysis. In this case, the associated regression parameters may be estimated with substantial uncertainty and the usual model comparison methods are often inadequate in determining any associated improvement in model fit. Again, the decision may be more domain-based than statistical. In the study considered here, results were reported with and without a covariate that was considered to be important for sports scientists, and a deliberate decision was made to avoid formal model comparison. These issues of data transformation and model comparison for small samples merit further research.

Here we consider instead an analogous scaled effect defined in terms of the relative difference between the pairs of measurements:

$$y_i = (\text{post}_i\text{-pre}_i)/\text{pre}_i. \quad (14)$$

For both the unscaled response given by Eq (13) and the relative response given by Eq (14), the list of posterior estimates of interest were:

- the differences between the two experimental training regimens (IHE, LHTL) and the Placebo group, given by δ_{12} and δ_{13}, respectively, and the difference between the two training regimens IHE and LHTL, given by δ_{23};
- Cohen's *d* for each of the two experimental regimens compared with the Placebo regimen, given by $d_{12} = \delta_{12}/\sigma_{12}$ for IHE and $d_{13} = \delta_3/\sigma_{13}$ for LHTL;
- Cohen's *d* for the standardized difference between LHTL versus IHE, given by $d_{23} = \delta_{23}/\sigma_{23}$;
- the probabilities that the standardized difference between the IHE training regimen and the Placebo exceed the 'smallest worthwhile change' (*SWC*, specified as a standardised change of *0.2* based on previous recommendations [18]), denoted by $SWCU_{12} = \Pr(d_{12} > 0.2)$ and $SWCL_{12} = \Pr(d_{12} < -0.2)$;
- analogous probabilistic comparisons with the *SWC* for the difference between the LHTL training regimen and the Placebo, and the LHTL and IHE training regimens,
- the posterior distributions of the expected outcome $\text{E}(y_{ij}) = \beta_0 + \beta_1 X + \beta_2 I_{j=1} + \beta_3 I_{j=2}$ for the *i*th individual under the *j*th training regimen (where $I_{j=1} = 1$ if the treatment is IHE and = 0 otherwise, and $I_{j=2} = 1$ if the treatment is LHTL and = 0 otherwise); the expected outcome, obtained by substituting the simulated parameter values ($\beta_0, \beta_1, \beta_2, \beta_3$) into this equation at each MCMC simulation,
- the analogous posterior predicted outcome for each individual under each training regimen, which allows for within-subject variation around the expected outcome, i.e., $y_{ij}^{\text{pred}} = y_{ij}^{\text{pred}} + e_{ij}, \quad e_{ij} \sim \text{N}(0, \sigma^2)$, which is obtained in the same manner as above,

- the ranks of the individuals based on their expected and predicted outcomes under each of the treatment regimens; again, this is a probability distribution, reflecting the fact that rankings may change depending on the precision of the estimated treatment effects and within-subject variation.

Note that although the denominator of the Cohen's d_{kl} values can be calculated using the traditional equation, i.e., $\sigma_{kl} = \sqrt{\text{Var}(\delta_l - \delta_k)} = \sqrt{((v_l \text{Var}(\delta_l) + v_k \text{Var}(d_k))/(v_l + v_k))}$, this can also be directly calculated using the simulated values of d_{kl} obtained from the MCMC iterations, i.e., $\sigma_{kl} = \sqrt{\text{Var}(d_{kl})}$.

Based on exploratory plots of the relationships between the observed pre- and post-training values of Hbmass, RunEcon and La-max among the three groups, and with the covariate, two analyses of the data were undertaken. In the first analysis, the covariate was excluded and the model was fit using Eqs (3) and (4). In the second analysis, the covariate was included given previous work showing that training load can influence the hemopoietic response [19] and the model was fit using Eq (11). The models were implemented using the statistical software R, with packages BRugs and R2WinBugs, which call WinBUGS [15, 20, 21], and MCMCregress in the MCMCpack library [22]. Estimates were based on 150,000 MCMC iterations, after discarding an initial burn-in of 50,000 iterations. For comparability with Humberstone-Gough *et al* [9], the results of the second analysis are reported below. The R code for this model is presented as a text file in S1 Text.

As described above, the primary analyses for the case study utilized an uninformative prior specification for β in Eq (12), which was obtained by setting the values of the prior mean vector b_0 and prior precision matrix B_0 to zero. The impact of informative priors was evaluated by considering a range of non-zero values for these terms, with Hbmass as the response measure. The values were motivated by the results of a recent meta-analysis of training regimens on Hbmass [23], which reported a mean response of 1.08% increase in Hbmass per 100 hours of LHTL training. Based on the study of Humberstone-Gough with 240 hours of exposure, the prior expectation is thus that the mean increases for the LHTL and IHE groups would be 2.6% and 0% respectively. The latter figure is also supported by a report that 3 h/day at 4000–5,500 m was inadequate to increase Hbmass at all [24]. This literature also provides a prior expectation of 0% increase in Hbmass of the Placebo group. The 2013 meta-analysis [23] also provided an estimate of 2.2% for the within-subject standard deviation of Hbmass.

Results

The distribution of the covariate X (representing the % change in weekly training load from pre- to post-camp) within and among the three training regimens (Placebo, IHE, LHTL) is displayed in Fig 1. The plots show that there is non-negligible variation between individuals within a regimen with respect to this variable and substantive differences between the regimens. It is clear that adjustment needs to be made for X before evaluating the comparative impact of the three regimens. This is accommodated in the regression model described in Eq (11).

Scatterplots of the unscaled differences given by Eq (13) and scaled differences given by Eq (14) are presented in Figs 2–4. Based on these plots, there is no clear visual association between the three measurements under consideration in this case study (Hbmass, RunEcon and La-max), or between these measurements and the covariate.

Plots of the posterior distributions of the differences between the training regimens, IHE vs Placebo, LHTL vs Placebo, LHTL vs IHE, given by δ_{12}, δ_{13} and δ_{23}, respectively, are shown in Fig 5. Corresponding posterior estimates of the effects (mean, s.d., 95% and 90% credible intervals) are given in Table 1.

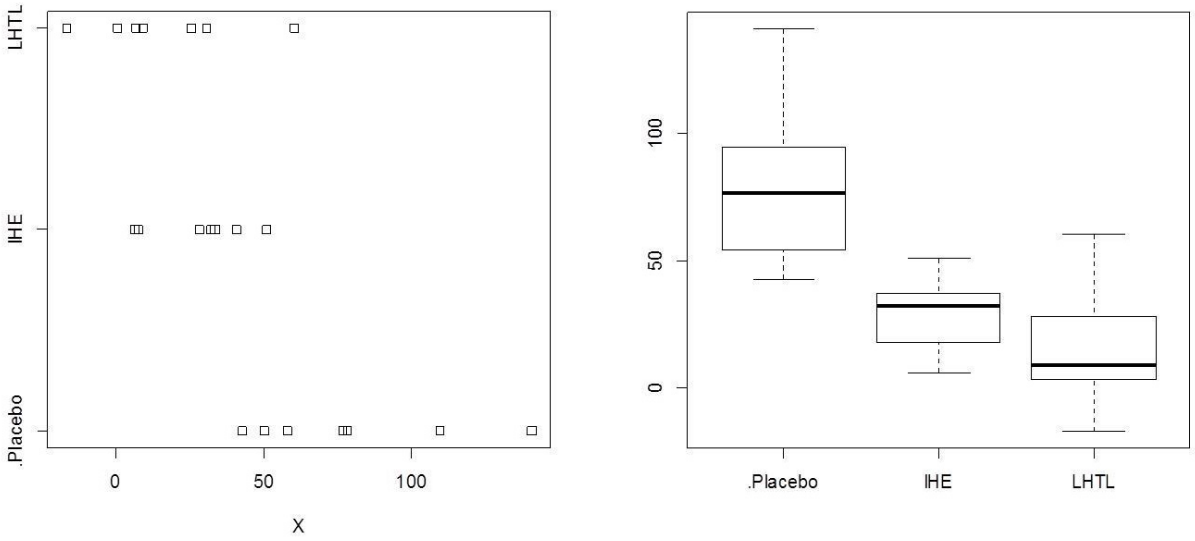

Fig 1. Exploratory analyses comprising stripcharts (left) and boxplots (right) for the covariate X in the three training regimens (Placebo, Intermittent Hypoxic Exposure (IHE), Live High Train Low (LHTL)), where X is a measure of the percent change in training load for each of the 23 individuals in the study. (See text for details.).

doi:10.1371/journal.pone.0147311.g001

From Fig 5 and Table 1, it can be seen that for Hbmass and RunEcon, although there is a slight detrimental effect of IHE and a slight beneficial effect of LHTL compared with the Placebo, these are not substantive: a difference of 0 is reasonably well supported by the posterior distributions. However, this slight differential in response results between IHE and LTHL: a difference of 0 appears to have less support in the posterior densities; the 90% credible interval does not include 0 and the posterior probability that Cohen's *d* exceeds the SWC is 0.96 and 0.93 for Hbmass and RunEcon respectively. These outcomes strongly indicate that LHTL is substantively better than IHE for both of these outcome measures.

In contrast, for La-max, both IHE and LHTL show a substantive beneficial effect compared with the Placebo, with the corresponding 95% (and hence 90%) credible intervals excluding 0

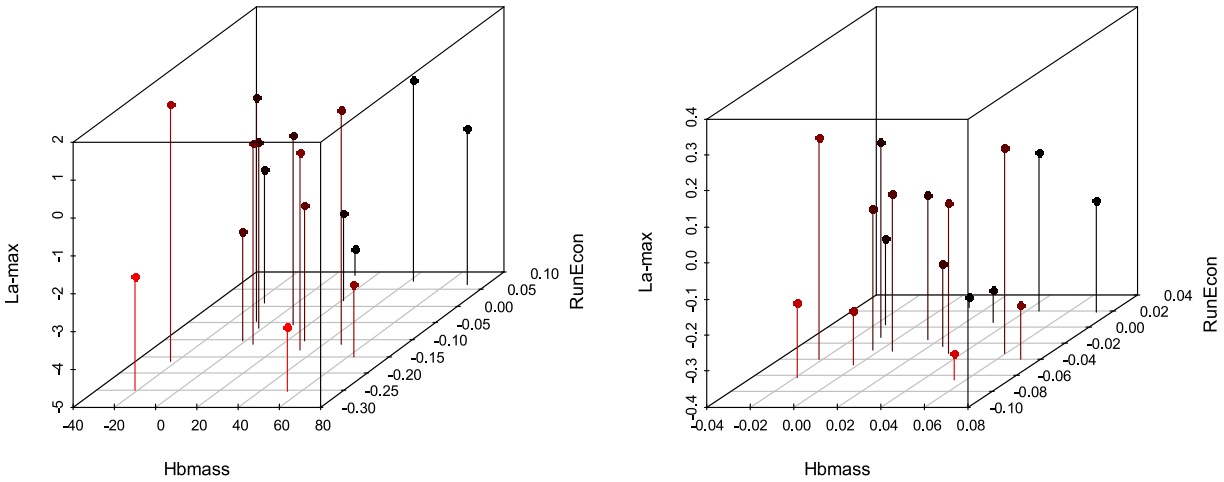

Fig 2. Three-dimensional scatterplot of the three measurements, Hemoglobin Mass (Hbmass), Running Economy (RunEcon) and maximum blood lactate concentration (La-max), unscaled data (left) and scaled data (right). Unscaled data are calculated as $post_i$–pre, and scaled data are calculated as ($post_i$–pre_i) / pre_i.

doi:10.1371/journal.pone.0147311.g002

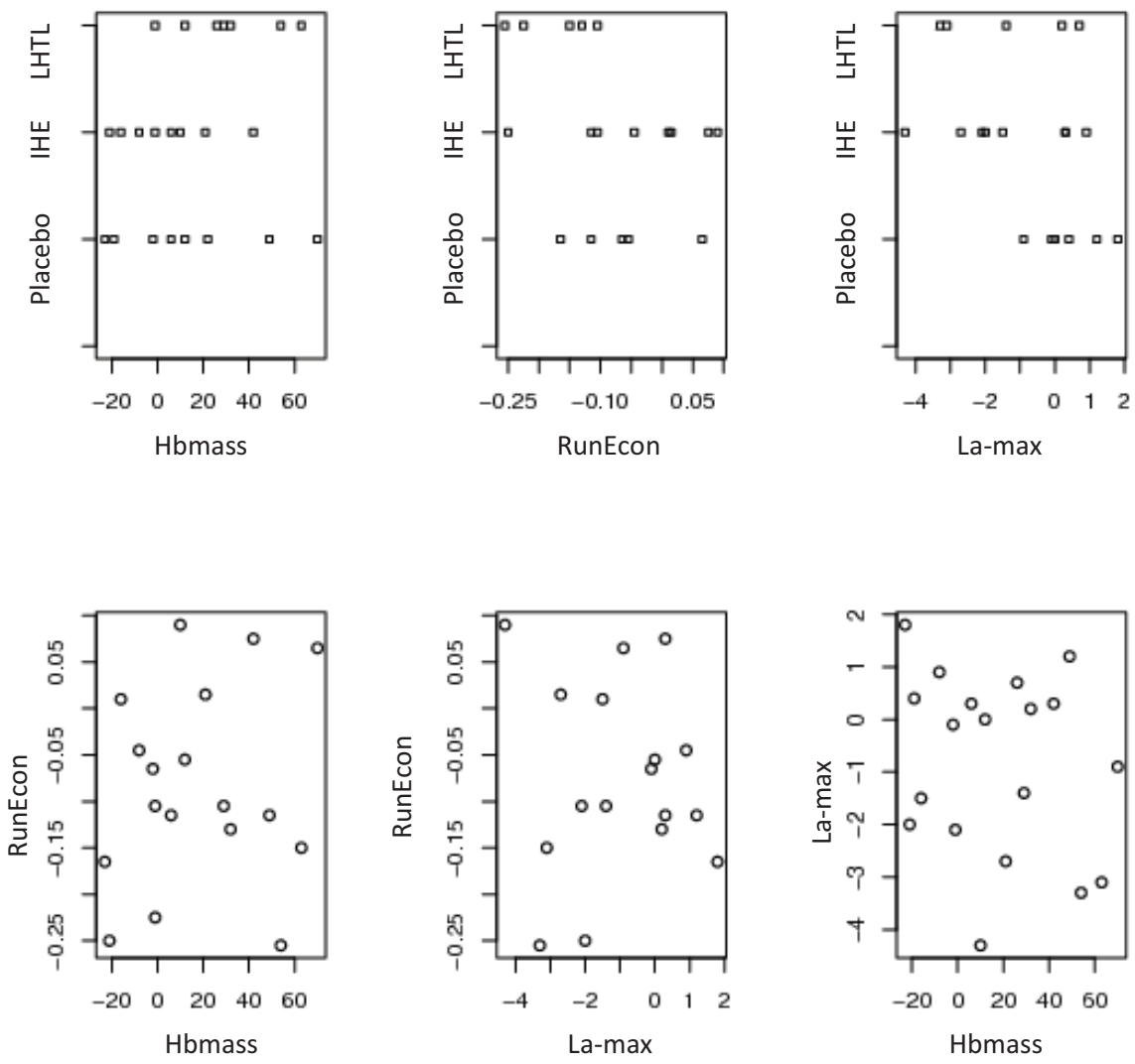

Fig 3. Two-dimensional scatterplots of the three measurements of Hbmass, RunEcon and La-max, under three regimes Placebo, Intermittent Hypoxic Exposure (IHE) and Live High Train Low (LHTL), unscaled data.

doi:10.1371/journal.pone.0147311.g003

and a probability of 0.97 that Cohen's *d* exceeds the SWC. As a consequence, the difference between LHTL and IHE is thus attenuated for this outcome measure.

Posterior estimates of parameters of interest for the scaled (relative) measures are shown in Fig 6 and Table 2. The figures and table confirm the above results. Similar to the unscaled effects, there is no clear visual association between two of the measurements under consideration in this case study (Hbmass and RunEcon), or between these measurements and the covariate of change in weekly training load. However, there is a clear difference in the values of the covariate among individuals in the Placebo group compared with the two training regimens (LHTL and IHE). The two training regimens both appear to substantively improve La-max, even after accounting for training-induced changes in the individual athletes. The direct probabilistic comparisons with the SWC provide more complete information about these treatments based on these data.

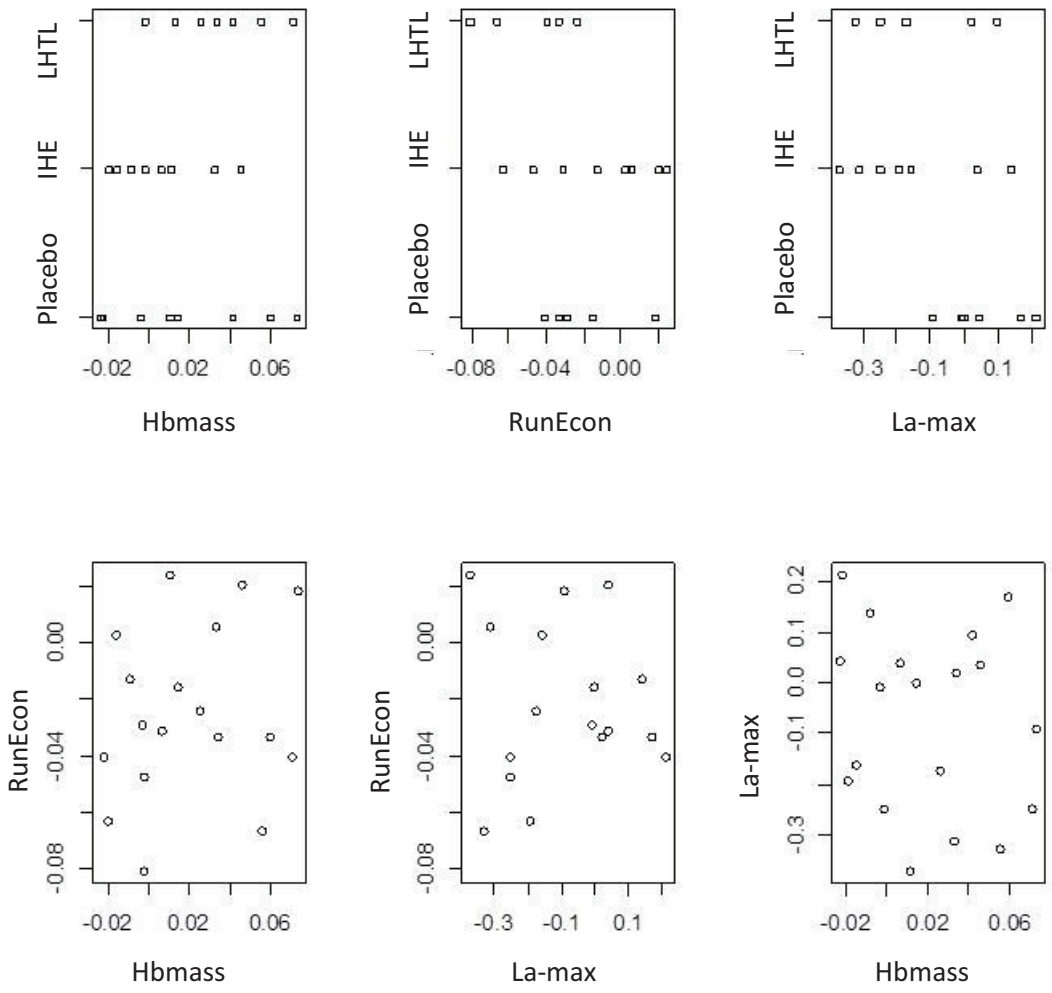

Fig 4. Two-dimensional scatterplots of the three measurements of Hbmass, RunEcon and La-max, under three regimes Placebo, Intermittent Hypoxic Exposure (IHE) and Live High Train Low (LHTL), scaled data.

doi:10.1371/journal.pone.0147311.g004

The posterior expected outcome of Hbmass for each individual under each regimen is illustrated in Fig 7, for the unscaled data. The boxplots indicate the distribution of possible outcomes, with the box corresponding to the middle 50% of values and the limits of the bars corresponding to the minimum and maximum values. The corresponding expected rank and associated interquartile range for the 23 individuals are reported in Table 3. It is noted that the predictions and ranks are substantively driven by the covariate values in this model, with comparatively much less influence from the effect of the training regimens. Hence Table 3 displays only a selection of results.

A comparison of two of the primary outcome measures Hbmass and RunEcon based on the Bayesian and magnitude-based inference approach is presented in Table 4. Note that the two sets of results differ slightly not only because of differences in analytic method, but also because of differences in modelling. For example, the magnitude-based inferences are based on a log-transformed response forecast to a covariate value (a 44% increase in weekly training load), with covariate adjustment undertaken within each treatment group; in contrast, the Bayesian inferences are based on the unadjusted and relative responses forecast to the mean covariate

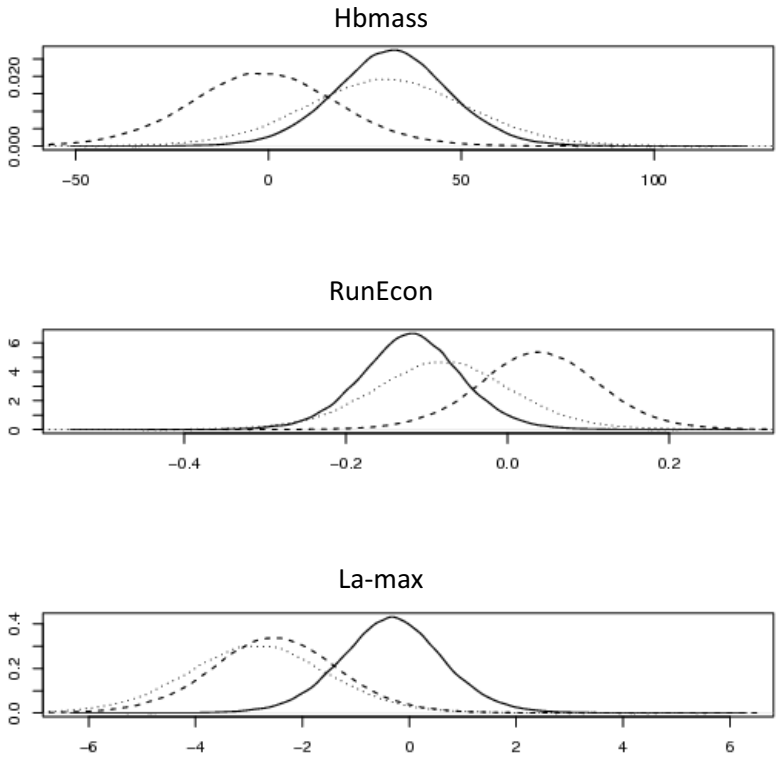

Fig 5. Posterior densities of the three measurements, Haemoglobin Mass, Running Economy and Running Maximum Lactate, comparing Live High Train Low (LHTL) vs Intermittent Hypoxic Exposure (IHE) (solid line), LHTL vs Placebo (dotted line) and IHE vs Placebo (dashed line), unscaled data.

doi:10.1371/journal.pone.0147311.g005

value and adjustment is undertaken using all of the data for reasons of small sample size. Furthermore, as described above, the method of computation of the denominator of the standardized values is not based on asymptotics in the Bayesian analysis, which makes a difference for small samples. Notwithstanding these differences, the overall conclusions are similar for the two sets of analyses. For Hbmass, the Bayesian analysis indicated a substantially higher increase for LHTL with both unscaled and scaled data, with magnitude-based analysis indicating possibly higher for LHTL with unscaled data, and likely higher with scaled data. Similarly for RunEcon the outcomes were similar between the analytical approaches—the Bayesian analysis indicated a substantial improvement (lower oxygen cost) with both unscaled and scaled data, while magnitude-based analysis indicated possibly lower oxygen cost in both cases.

Comparison of the expected values of Hbmass and La-max under each of the training regimens is further illustrated in Fig 8. The diagonal line indicates no treatment effect. The cloud of points represents the values obtained from the MCMC simulations in the Bayesian analysis. Displacement of the cloud from the line indicates that that there is an expected improvement

Table 1. Posterior estimates based on unscaled data.

Hbmass				
Posterior parameter estimates (units of grams)				
Effect	Mean	s.d.	95% CI	90% CI
X	0.25	0.25	-0.25, 0.74	-0.17, 0.66
IHE	-1.4	19.8	-40.5, 37.8	-33.8, 30.9
LHTL	30.7	21.7	-12.4, 73.4	-4.9, 66.2
LHTL-IHE	32.0	15.3	1.9, 62.2	7.04, 57.2
Cohen's d				
Effect	Mean	s.d.	95% CI	90% CI
IHE	-0.07	1.0	-2.1, 1.9	-1.7. 1.6
LHTL	1.4	1.0	-0.57, 3.4	-0.23, 3.0
LHTL-IHE	2.1	1.0	0.12, 4.1	0.46, 3.7
Prob. Cohen's d <> 0.2				
Parameter	Prob. *d*<-0.2	Prob. *d*>0.2		
IHE	0.45	0.39		
LHTL	0.052	0.89		
LHTL-IHE	0.013	0.97		
RunEcon				
Posterior parameter estimates (unit of L/min)				
Effect	Mean	s.d.	95% CI	90% CI
X	0.00045	0.0010	-0.0016, 0.0025	-0.0012, 0.0021
IHE	0.039	0.079	-0.12, 0.20	-0.09, 0.17
LHTL	-0.080	0.090	-0.26, 0.097	-0.23, 0.065
LHTL-IHE	-0.12	0.064	-0.25, 0.0094	-0.22, -0.014
Cohen's d				
Effect	Mean	s.d.	95% CI	90% CI
IHE	0.50	1.0	-1.5, 2.5	-1.1, 2.1
LHTL	-0.89	1.0	-2.9, 1.1	-2.5, 0.73
LHTL-IHE	-1.85	1.0	-3.8, 0.15	-3.5, -0.22
Prob. Cohen's d <> 0.2				
Parameter	Prob. *d*<-0.2	Prob. *d*>0.2		
IHE	0.23	0.62		
LHTL	0.77	0.13		
LHTL-IHE	0.95	0.023		
La-max				
Posterior parameter estimates (units of mmol/L)				
Effect	Mean	s.d.	95% CI	90% CI
X	-0.018	0.015	-0.050, 0.013	-0.044, 0.0076
IHE	-2.5	1.3	-5.0, -0.06	-4.6, -0.50
LHTL	-2.8	1.3	-5.6, -0.10	-5.10-, -0.59
LHTL-IHE	-0.32	1.0	-2.3, 1.66	-1.2, 1.3
Cohen's d				
Effect	Mean	s.d.	95% CI	90% CI
IHE	-2.0	1.0	-4.0, -0.054	-3.7, -0.40
LHTL	-2.1	1.0	-4.0, -0.070	-3.7, -0.48
LHTL-IHE	-0.32	1.0	-2.3, 1.7	-2.0, 1.3

(*Continued*)

Table 1. (*Continued*)

Prob. Cohen's d <> 0.2		
Parameter	Prob. *d*<-0.2	Prob. *d*>0.2
IHE	0.97	0.015
LHTL	0.98	0.015
HTL-IHE	0.55	0.29

doi:10.1371/journal.pone.0147311.t001

or decline in the outcome measure associated with the respective treatment, and the range of values for which this is anticipated to take effect.

The alternative priors that were motivated by the available external information are shown in Table 5. The consequent changes in the parameter values arising from the incorporation of these priors in the model are also shown in this table. It is clear that although the parameter estimates change slightly, the inferences reported above are generally robust to relatively small changes in the priors. However, the posterior estimates start to differ in a natural manner when the priors become more informative with respect to either the mean or variance. It is also noted that, reassuringly, the original (vague prior) setting yielded a posterior estimate of a relative increase of 2.6% in Hemoglobin mass under the LHTL regimen, which is equivalent to the anticipated value based on the (independent) prior information.

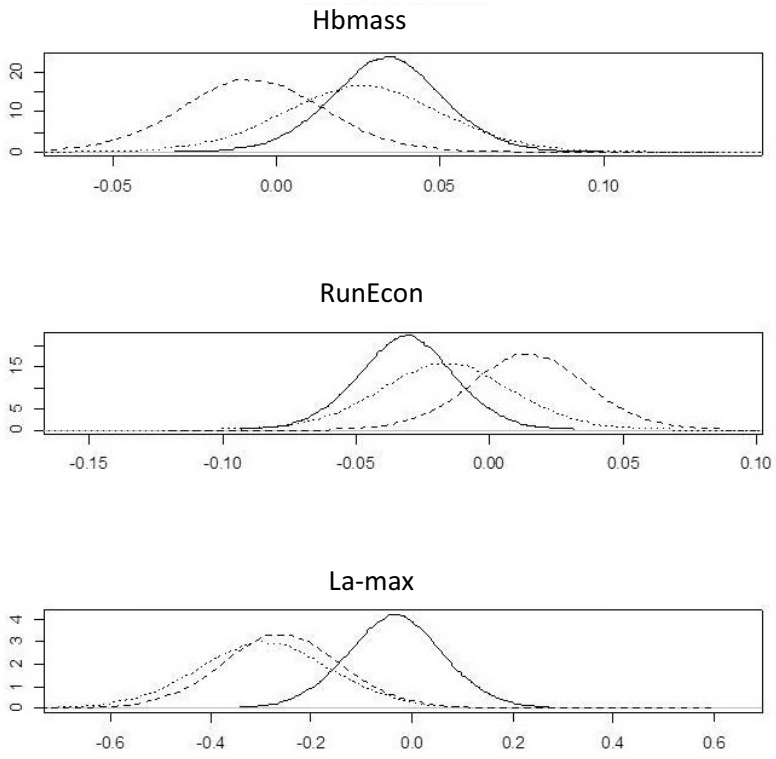

Fig 6. Posterior densities of the three measurements, Haemoglobin Mass, Running Economy and Running Maximum Lactate, comparing Live High Train Low (LHTL) vs Intermittent Hypoxic Exposure (IHE) (solid line), LHTL vs Placebo (dotted line) and IHE vs Placebo (dashed line), scaled data.

doi:10.1371/journal.pone.0147311.g006

Table 2. Posterior estimates based on scaled data.

Hbmass				
Posterior parameter estimates (units of percent / 100)				
Effect	Mean	s.d.	95% CI	90% CI
X	0.00020	0.00029	-0.00038, 0.00077	-0.00028, 0.00067
IHE	-0.0075	0.023	-0.053, 0.038	-0.045, 0.030
LHTL	0.026	0.025	-0.023, 0.076	-0.015, 0.068
LHTL-IHE	0.034	0.018	-0.0011, 0.069	0.0050, 0.063
Cohen's d				
Effect	Mean	s.d.	95% CI	90% CI
IHE	-0.33	1.0	-2.3, 1.7	-1.2, 1.3
LHTL	1.1	1.0	-0.93, 3.0	-0.60, 2.7
LHTL-IHE	1.9	1.0	-0.059, 3.9	0.28, 3.6
Prob. Cohen's d <> 0.2				
Parameter	Prob. *d*<-0.2	Prob. *d*>0.2		
IHE	0.55	0.29		
LHTL	0.10	0.81		
LHTL-IHE	0.019	0.96		
RunEcon				
Posterior parameter estimates (units of percent / 100)				
Effect	Mean	s.d.	95% CI	90% CI
X	0.00023	0.00031	-0.00038, 0.00083	-0.00027, 0.00072
IHE	0.015	0.023	-0.032, 0.061	-0.024, 0.053
LHTL	-0.016	0.027	-0.069, 0.037	-0.060, 0.027
LHTL-IHE	-0.031	0.019	-0.069, 0.0071	-0.062, 0.00016
Cohen's d				
Effect	Mean	s.d.	95% CI	90% CI
IHE	0.63	1.00	-1.4, 2.6	-1.0, 2.3
LHTL	-0.61	1.00	-2.6, 1.4	-2.2, 1.0
LHTL-IHE	-1.62	1.00	-3.6, 0.37	-3.3, 0.0085
Prob. Cohen's d <> 0.2				
Parameter	Prob. *d*<-0.2	Prob. *d*>0.2		
IHE	0.19	0.68		
LHTL	0.67	0.20		
LHTL-IHE	0.93	0.035		
La-max				
Posterior parameter estimates (units of percent / 100)				
Effect	Mean	s.d.	95% CI	90% CI
X	-0.0019	0.0016	-0.0051, 0.0013	-0.0045, 0.00072
IHE	-0.26	0.13	-0.51, -0.0094	-0.47, -0.054
LHTL	-0.29	0.14	-0.58, -0.014	-0.52, -0.065
LHTL-IHE	-0.034	0.10	-0.24, 0.17	-0.20, 0.13
Cohen's d				
Effect	Mean	s.d.	95% CI	90% CI
IHE	-2.6	1.0	-4.0, -0.074	-3.7, -0.43
LHTL	-2.1	1.0	-4.1, -0.10	-3.7, -0.46
LHTL-IHE	-0.33	1.0	-2.3, 1.7	-2.0, 1.3

(*Continued*)

Table 2. (*Continued*)

Prob. Cohen's d <> 0.2		
Parameter	Prob. *d*<-0.2	Prob. *d*>0.2
IHE	0.97	0.014
LHTL	0.97	0.014
LHTL-IHE	0.56	0.29

doi:10.1371/journal.pone.0147311.t002

Discussion

In 2008, Barker and Schofield [7] suggested that "to correctly adopt the type of inference advocated by Batterham and Hopkins [6], sport scientists need to use fully Bayesian methods of analysis". They also noted that most sport scientists are not trained in Bayesian methods, likely because this approach has only become commonplace as a statistical technique in approximately the last 20 years. To help make the Bayesian approach more accessible for those working in exercise science and sports medicine, we have provided here both a worked example (using statistical software) together with a description of the underlying models. We hope that this template will encourage those who deal with small samples and small effects to explore the full Bayesian method, which is well suited to the analysis of small samples. Other supporting information, where available, can be represented via the prior and hence formally and transparently incorporated with the data. In the absence of such information, the uncertainty induced by small samples is properly incorporated in the posterior estimates and inferences. In both of these situations, the analytical decision-making is enhanced, in support of the ultimate practical/clinical decision-making undertaken by sports practitioners.

Case study re-interpreted with Bayesian inferences

An experimental study by Humberstone-Gough and colleagues reported changes (mean ± 90% confidence interval) in Hbmass of -1.4 ± 4.5% for IHE compared with Placebo and 3.2 ± 4.8%

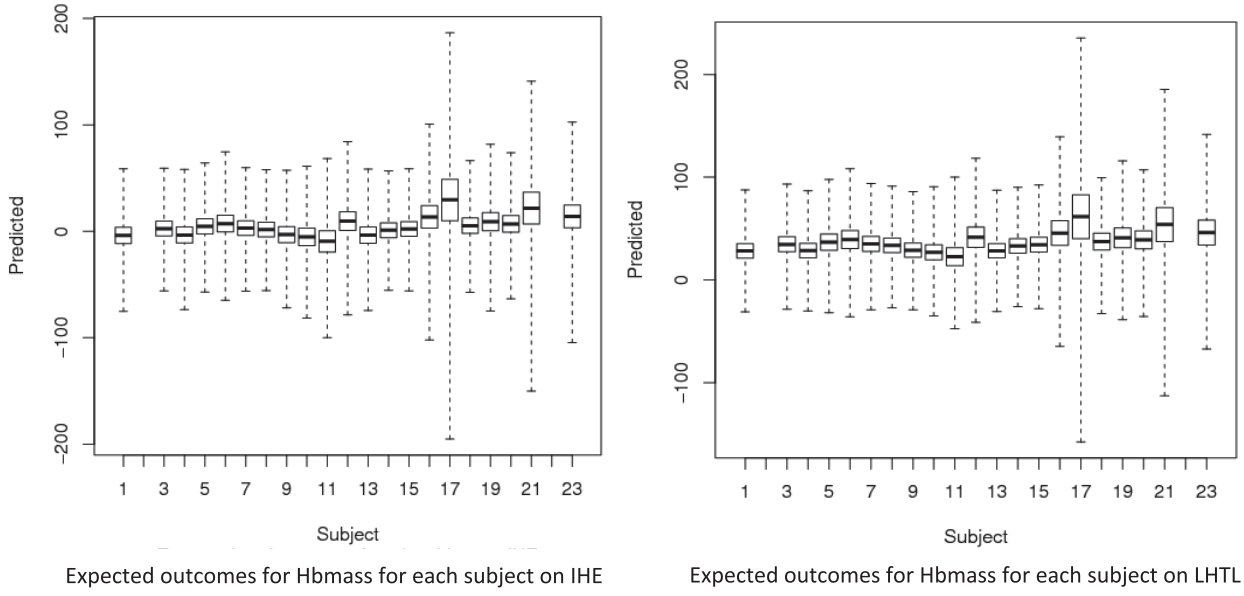

Fig 7. Boxplots of the posterior expected outcomes for Hbmass for each individual in the study, under each of the two training regimens Intermittent Hypoxic Exposure (left) and Live High Train Low (right).

doi:10.1371/journal.pone.0147311.g007

Table 3. Expected rank and associated interquartile range for the 23 individuals in the study.

ID	Hbmass		RunEcon		La-max	
	Mean	IQR	Mean	IQR	Mean	IQR
1	3	3–19	3	3–19	19	3–19
2	NA	NA	NA	NA	NA	NA
3	10	10–12	10	10–12	12	10–12
4	5	5–17	5	5–17	17	5–17
5	12	10–12	12	10–12	10	10–12
6	15	7–15	15	7–15	7	7–15
7	11	11–11	11	11–11	11	11–11
8	8	8–14	8	8–14	14	8–14
9	6	6–16	6	6–16	16	6–16
10	2	2–20	2	2–20	20	2–20
11	1	1–21	1	1–21	21	1–21
12	17	5–17	17	5–17	5	5–17
13	4	4–18	4	4–18	18	4–18
14	7	7–15	7	7–15	15	7–15
15	9	9–13	9	9–13	13	9–13
16	18	4–18	18	4–18	4	4–18
17	21	1–21	21	1–21	1	1–21
18	13	9–13	13	9–13	9	9–13
19	16	6–16	16	6–16	6	6–16
20	14	8–14	14	8–14	8	8–14
21	20	2–20	20	2–20	2	2–20
22	NA	NA	NA	NA	NA	NA
23	19	3–19	19	3–19	3	3–19

doi:10.1371/journal.pone.0147311.t003

for LHTL compared with Placebo [9]. For RunEcon the authors reported 'no beneficial changes' for IHE compared with Placebo, and a change of 2.8 ± 4.4% for LHTL compared with Placebo. Although the analyses were undertaken using different outcome measures and a slightly different analytical model, the conclusions based on the posterior estimates and probabilities obtained from the Bayesian analysis reported above are broadly consistent with those reported by Humberstone-Gough *et al*. Importantly, the Bayesian approach allows a much more direct probabilistic interpretation of credible intervals and posterior probabilities; for example, the probability that the mean change in Hbmass after LHTL compared with the change after IHE is greater than the smallest worthwhile change (0.2) is 0.96.

Cohen's effect size magnitudes are well established [11] but the selection of a small effect ($d = 0.2$) as the threshold value for a worthwhile change or difference has been questioned. In the sporting context, worthwhile changes in competition performance, which can alter medal rankings, have been derived [25] as approximately 0.3 times the within-subject standard deviation [26, 27], or ~0.3–1% of performance time in a range of sports [28–30]. Empirical evidence confirms that small effects (on competitive performance) are worthwhile for elite athletes and of practical relevance for coaches and scientists attempting to understand the likely benefit or harm of training regimen, lifestyle intervention or change in technique. The full Bayesian approach provides a robust and acceptable method of estimating the likelihood of a small effect. For instance, in the Humberstone-Gough et al. case study Hbmass increased ~21 g (or by 2.3%) more in LHTL vs IHE. Given that every gram of hemoglobin can carry ~4 mL O_2, [31], it is

Table 4. Analysis of pre- to post-training measurements for LHTL vs IHE–outcomes for Bayesian and Magnitude-based Inferences for both unscaled and scaled data. SD = standard deviation, CL = confidence limits, CI = credible interval.

Analysis	Measure	Hemoglobin Mass (g)	Running Economy ($L.min^{-1}$)
Bayesian Unscaled	Mean ± SD	21 ± 17	-0.17 ± 0.052
	90% CI	-6, 48	-0.25, -0.08
	Cohen's d; 90% CI	1.26; -0.37, 2.90	-3.20; -4.84, -1.57
	Probability |d|>0.2	0.931	0.998
	Qualitative inference	Higher	Lower
Magnitude-based Inference	Mean; 90% CL	36; -5, 78	-0.13; -0.22, 0.04
	Cohen's d; 90% CL	0.18; -0.02, 0.39	-0.20; -0.34, -0.07
	Qualitative inference	Possibly Higher	Possibly Lower
Bayesian Scaled	Mean ± SD (% / 100)	0.023 ± 0.019	-0.042 ± 0.017
	90% CI	-0.008, 0.054	-0.069, -0.015
	Cohen's d; 90% CI	1.21; -0.42, 2.85	-2.51; -4.14, -0.88
	Probabilty |d|>0.2	0.926	0.993
	Qualitative inference	Higher	Lower
Magnitude-based Inference	Smallest worthwhile difference (% / 100)	0.016	0.019
	Difference ± SD	0.047 ± 0.035	-0.028 ± 0.044
	Cohen's d; 90% CL	0.20; 0.05, 0.35	-0.14; -0.34, 0.07
	Qualitative inference	Likely Higher	Possibly Lower

doi:10.1371/journal.pone.0147311.t004

reasonable to infer that this small increase in Hbmass is likely beneficial to overall oxygen transport capacity. The corresponding 95% credible interval for this comparison of absolute change ranged from -11.8 to +53.8 g, but on balance the probability is >0.8 that the true increase in Hbmass is substantial (worthwhile), which should be sufficient encouragement for most

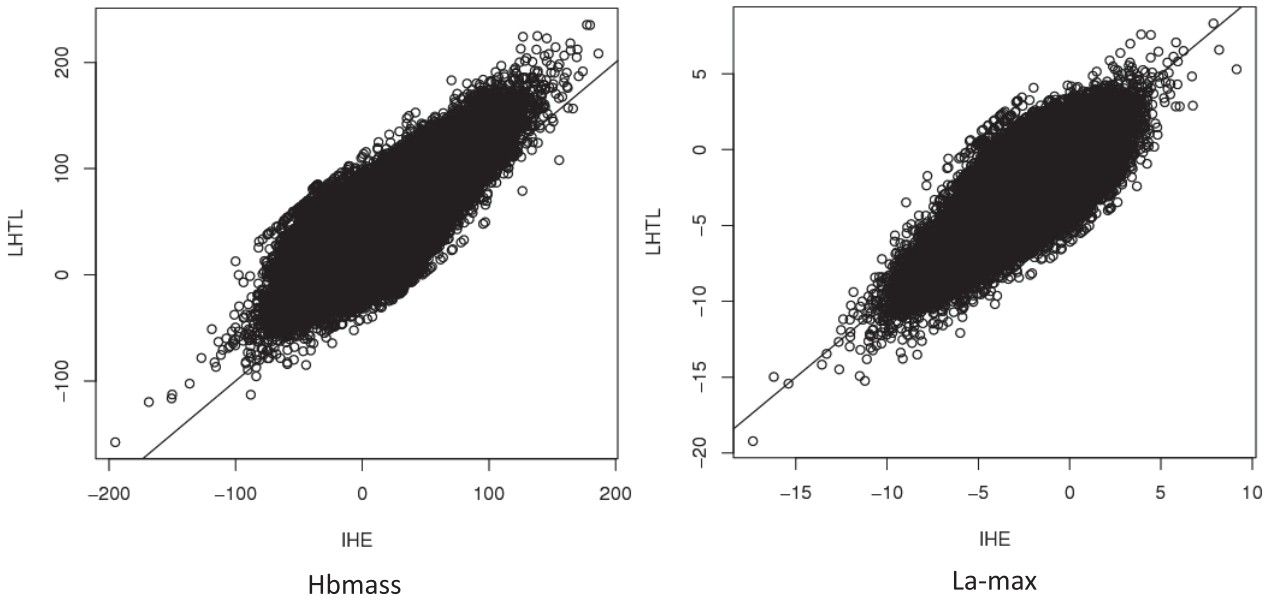

Fig 8. Comparison of the posterior distributions of the expected measurements of Hbmass (left) and La-max (right) under each of the training regimens Intermittent Hypoxic Exposure (IHE) and Live High Train Low (LHTL), unscaled data.

doi:10.1371/journal.pone.0147311.g008

Table 5. Configurations of hyperparameter values for informative priors in the Bayesian model [Eqs (11 and 12)]. Here b_0 and B_0 denote respectively the prior mean vector and precision matrix for the regression coefficients, and $c_0/2$ and $d_0/2$ denote respectively the shape parameter and scale parameter for the inverse Gamma prior on σ^2 (the variance of the disturbances). These latter two parameters can be respectively interpreted as indicating the amount of information, and the sum of squared errors, from c_0 pseudo-observations, for the inverse Gamma prior on σ^2 (the variance of the residuals) [16]. Note that (a) depicts the baseline uninformative priors used in the primary analyses, whereas (b) to (h) illustrate seven alternate priors.

Setting	(a)	(b)	(c)	(d)	(e)	(f)	(g)	(h)
b_0	(0,0,0,0)	(0,0,0,2.6)	(0,0,0,2.6)	(0,0,0,2.6)	(0,0,0,0)	(0,0,0,0)	(0,0,0,0)	(0,0,0,2.6)
diag(B_0)	(0,0,0,0)	(0,0,.2,.2)	(5,5,5,5)	(0,0,5,5)	(0,0,0,0)	(0,0,0,0)	(0,0,5,5)	(0,0,0,0)
c_0	0.0001	0.0001	0.0001	0.0001	20	20	20	20
d_0	0.0001	0.0001	0.0001	0.0001	100	5	100	100
Int.	0.00047	0.0044	-0.0027	-0.0028	0.011	0.0060	-0.64	0.0060
	(0.026)	(0.026)	(0.026)	(0.026)	(1.4)	(0.30)	(0.28)	(0.30)
X	0.00020	0.00020	0.00026	0.00026	0.00018	0.00019	0.0062	0.00019
	(0.00029)	(0.00029)	(0.00029)	(0.00029)	(0.015)	(0.0034)	(0.0033)	(0.0034)
IHE	-0.0075	-0.0073	-0.0021	-0.0021	-0.017	-0.0096	0.41	-0.0096
	(0.023)	(0.023)	(0.023)	(0.023)	(1.2)	(0.27)	(0.23)	(0.27)
LHTL	0.026	0.027	0.035	0.035	0.024	0.026	0.79	0.26
	(0.025)	(0.025)	(0.025)	(0.025)	(1.3)	(0.29)	(0.27)	(0.29)
σ^2	0.0011	0.0011	0.0011	0.0011	2.9	0.14	0.17	0.14
	(0.00042)	(0.00042)	(0.00043)	(0.00043)	(0.71)	(0.035)	(0.047)	(0.045)

doi:10.1371/journal.pone.0147311.t005

scientists and coaches to utilize altitude training to increase Hbmass–a position also supported by a meta-analysis of Hbmass and altitude training [23]. Likewise in the Humberstone-Gough et al. case study RunEcon improved (was lower) by ~0.17 $L.min^{-1}$ (or lower by 4.2%) more in LHTL vs IHE. The associated 95% credible interval for this comparison of relative change ranged from –0.9 to -7.5%, with a probability of ~0.99 that the true decrease in submaximal oxygen consumption is substantial (worthwhile). Although contentious [32], an improved running economy after altitude training is advantageous to distance running performance because it reduces the utilization of oxygen at any given steady-state running speed [33, 34].

Limitations of quasi-Bayesian approaches

Batterham and Hopkins (2006) have challenged the frequentist approach as being too conservative, and provided a useful, if somewhat unconventional, framework for interpreting small effects. The so-called magnitude-based approach emerging in sports science [18, 26, 35] is based on defining and justifying clinically, practically or mechanistically meaningful values of an effect. Confidence intervals are then used to interpret uncertainty in the effect in relation to these reference or threshold values. Much discussion has centred on the legitimacy of using vague priors in the magnitude-based approach and whether prior knowledge is actually useful in all cases [36]. There are inferential limitations to their approach [7, 8] which can be circumvented by using a full Bayesian approach that we have elaborated here.

A major criticism of the approach suggested by Batterham and Hopkins (2006) is that, contrary to the authors' claims, their method is not (even approximately) Bayesian and that a Bayesian formulation of their approach would indeed make prior assumptions about the distribution of the true parameter values. Barker and Schofield (2008) suggest that the underlying prior distribution would be uniform, which makes a clear assumption about the parameter values (that any parameter value in the defined range is equally likely) and which can be influenced by transformations of the parameter. As demonstrated in our paper, a Bayesian formulation of the problem considered by Batterham and Hopkins (2006) can quite easily be

constructed, using a reference prior which is arguably vague (often referred to as the Jeffreys prior [10]). Moreover, there are clear and natural links between the frequentist distributions based on sampling theory and the Bayesian posterior distributions under these prior assumptions. The use of the reference prior for the estimation and comparison problem considered in this paper is well-founded, theoretically sound and very commonly employed [10]. As discussed in the Methods section, however, other priors can also be considered, particularly if there is other information available to complement the analysis.

Another criticism levelled at Batterham and Hopkins (2006) by Barker and Schofield (2008) is their choice and use of an expanded set of categories, based on a non-standard choice of the thresholds used to define the categories, the use of different thresholds for different problems (e.g., sometimes 0.025 and 0.975 instead of 0.05 and 0.95), and the descriptors used to label the categories, namely 'almost certainly not,. . .almost certainly'). However, while the expanded set of categories proposed by Batterham and Hopkins is not 'standard' in classical statistics, this does not mean that it is wrong, misleading or not useful. Indeed, such categorizations can be very useful *if* they are clearly justified, interpreted properly and provide additional decision support for clinical (or, in this case, sporting) interventions. Even in 'traditional' statistics, some statisticians suggest that a p-value less than 0.10 indicates 'substantive' evidence against the null hypothesis, while other statisticians would not counsel this. Similarly, although a p-value of 0.05 is almost overwhelmingly taken as the 'significance level', many statisticians strongly advise against its unconsidered use and suggest that other levels (such as 0.01 or 0.10) may be more appropriate for certain problems and desired inferences. A number of commentators in the sports science field have made similar observations [1, 37, 38]. The overwhelming advice is that the probabilities obtained as a result of statistical analysis must be useful in providing decision support for the problem at hand, and different probabilities can indeed be used if they are well justified, transparently reported and correctly interpreted.

The technical interpretation of a (frequentist) confidence interval is poorly understood by many practitioners. This has caused, and will continue to lead to, clumsy statements about the inferences that can be made on its basis. In contrast, an analogous Bayesian interval is directly interpretable: for example, a 95% credible interval indicates that the true parameter lies within this interval with an estimated probability of 0.95. Moreover, the analysis can be used to obtain other decision support statements such as a set of meaningful probabilities; for example, as demonstrated in the case study, one can obtain the probability that a particular parameter exceeds an objectively-derived threshold of clinical/practical/sporting interest. Of course, the particular decisions that are made on the basis of these probabilities remain the prerogative of the decision-maker. For example, the outcome of an intervention to improve athletic performance (e.g. a new experimental therapeutic treatment) may be classified as 'possible' in some cases (acceptable probability of improving performance, within minimal adverse effects, low cost, readily available, and legal in terms of anti-doping regulations), and hence lead to a decision of using, whereas in another context it may be deemed too risky (unacceptable risk of impairing performance, adverse effects on health and well-being, high cost and limited availability, and some uncertainty in meeting anti-doping regulations) and lead to no action. In practice, these decisions may not coincide with the traditional statement of a statistically significant effect at a 5% level [36]. In both cases, however, the decisions are enhanced by the richer probabilistic and inferential capability afforded by the Bayesian analysis.

In the context of small samples such as those encountered in this study, it is important to understand the nature and implications of the statistical assumptions underlying the adopted models and inferences. For example, in a standard linear regression model a common assumption is that the residuals (the differences between the observed and predicted values) are normally distributed. Note that this only applies to the residuals, not the explanatory or response

variables. This assumption was also adopted in the model and analysis presented in this paper. There is a rich literature about the appropriateness of this assumption for small sample sizes. Importantly, if the residuals are indeed normally distributed then the regression estimates will possess all three desirable statistical characteristics of unbiasedness, consistency, and efficiency among all unbiased estimators; however, even if they are not normally distributed they will still be unbiased (accurate) and consistent (improve with increasing sample size) but will only be most efficient (i.e. have smallest variance) among a smaller class of (linear unbiased) estimators [39]. The most obvious implication of non-normal residuals is that the inferences may not be as sharp, but by virtue of the central limit theorem the sampling distribution of the coefficients will approach a normal distribution as the sample size increases, under mild conditions. In our study, this was achieved by employing a single residual term across all groups which effectively increased the sample size. Feasible alternatives would have been to allow different residual variances for each group or to employ a robust regression approach, for example using a *t* distribution for the errors. It is also noted that the Bayesian estimates avoid some of the inferential concerns, since the credible intervals and probabilistic rankings are obtained from the MCMC samples, i.e., from the posterior distributions themselves, as opposed to relying on stronger asymptotic assumptions that are required for frequentist inferences

Another topical issue that has substantive implications for small sample analysis is reproducibility [40]. Indeed, the very measure of reproducibility arguably faces similar challenges as those reported here, and a Bayesian approach is arguably preferable over measures based on p-values or confidence intervals [41–43]. See also a recent blog article that discusses this topic (http://alexanderetz.com/2015/08/30/the-bayesian-reproducibility-project/). The current debates are often conducted in the context of large samples, so the challenge is much greater for studies such as the one presented here. This is another topic for future research.

Conclusion

We have demonstrated that a Bayesian analysis can be undertaken for small scale athlete studies and can yield comparable, but more directly interpretable and theoretically justified probabilistic outcomes compared with the so-called magnitude-based (quasi-Bayesian) approach. The model described here is one of the simplest Bayesian formulations, and can be expanded as needed to address other issues. Analytical approaches for small sample studies using full Bayesian, quasi-Bayesian, and frequentist decisions must be well justified, reported transparently and interpreted correctly.

Supporting Information

S1 Table. Data used in the case study.
(DOCX)

S1 Text. R code used in the analysis of the case study.
(DOCX)

Author Contributions

Conceived and designed the experiments: KM CD CG. Performed the experiments: KM CD. Analyzed the data: KM CD CR DP CG. Contributed reagents/materials/analysis tools: KM CD. Wrote the paper: KM CD CR DP CG.

References

1. Atkinson G, Batterham AM, Hopkins WG. Sports performance research under the spotlight. Int J Sports Med. 2012; 33: 949. doi: 10.1055/s-0032-1327755 PMID: 23165647
2. Ploutz-Snyder RJ, Fiedler J, Feiveson AH. Justifying small-n research in scientifically amazing settings: challenging the notion that only "big-n" studies are worthwhile. J Appl Physiol. 2014; 116: 1251–2. doi: 10.1152/japplphysiol.01335.2013 PMID: 24408991
3. Bacchetti P. Current sample size conventions: flaws, harms, and alternatives. BMC Medicine. 2010; 8: 17. doi: 10.1186/1741-7015-8-17 PMID: 20307281
4. Bacchetti P, Deeks SG, McCune JM. Breaking free of sample size dogma to perform innovative translational research. Sci Transl Med. 2011; 3(87): 87sp24.
5. Beck TW. The importance of a priori sample size estimation in strength and conditioning research. J Str Cond Res. 2013; 27: 2323–37.
6. Batterham AM, Hopkins WG. Making meaningful inferences about magnitudes. Int J Sports Physiol Perf. 2006; 1: 50–7.
7. Barker RJ, Schofield MR. Inference about magnitudes of effects. Int J Sports Physiol Perf. 2008; 3: 547–57.
8. Welsh AH, Knight EJ. "Magnitude-Based Inference": A Statistical Review. Med Sci Sports Exerc. 2014; 47: 874–84.
9. Humberstone-Gough C, Saunders PU, Bonetti DL, Stephens S, Bullock N, Anson JM et al. Comparison of live high: train low altitude and intermittent hypoxic exposure. J Sports Sci Med. 2013; 12: 394–401. PMID: 24149143
10. Gelman A, Carlin J, Stern H, Dunson D, Vehtari A, Rubin D. *Bayesian Data Analysis*. 3rd ed.: Chapman and Hall.; 2013, 64–9 p.
11. Cohen J. *Statistical power analysis for the behavioral sciences*. Hillsdale, New Jersey: Lawrence Erlbaum Associates; 1988, 1–17 p.
12. Gelman A, Carlin J, Stern H, Dunson D, Vehtari A, Rubin D. *Bayesian Data Analysis*. 3rd ed.: Chapman and Hall; 2013, 275–92 p.
13. Geman S, Geman D. Stochastic relaxation, Gibbs distributions and Bayesian restoration of images. IEEE Transactions on Pattern Analysis and Machine Intelligence. 1984; 6: 721–41. PMID: 22499653
14. Hedges LV, Olkin I. *Statistical Methods for Meta-Analysis*. Orlando: Academic Press; 1985.
15. Lunn DJ, Thomas A, Best N, Spiegelhalter D. WinBugs—a Bayesian modelling framework: concepts, structure and extensibility. Stats Computing. 2000; 10: 325–37.
16. Marin J, Robert C. *Bayesian Core*: *A Practical Approach to Computational Bayesian Statistics*. Springer; 2007.
17. Marin J, Robert C. *Bayesian Essentials with R*. Springer; 2014.
18. Hopkins WG, Marshall SW, Batterham AM, Hanin J. Progressive statistics for studies in sports medicine and exercise science. Med Sci Sports Exerc. 2009; 41: 3–13. doi: 10.1249/MSS.0b013e31818cb278 PMID: 19092709
19. Garvican LA, Martin DT, McDonald W, Gore CJ. Seasonal variation of haemoglobin mass in internationally competitive female road cyclists. Eur J Appl Physiol. 2010; 109: 221–31. doi: 10.1007/s00421-009-1349-2 PMID: 20058020
20. Sturtz S, Liggues U, Gelman A. R2WinBUGS: A package for running WinBUGS from R. J Stats Software. 2005; 12: 1–16.
21. Thomas A, O'Hara B, Ligges U, Sturtz S. Making BUGS Open. R News. 2006; 6: 12–7.
22. Martin A, Quinn K, Park J-H. MCMCpack: Markov chain Monte Carlo in R. J Stats Software. 2011; 429: 1–21.
23. Gore CJ, Sharpe K, Garvican-Lewis LA, Saunders PU, Humberstone CE, Robertson EY et al. Altitude training and haemoglobin mass from the optimised carbon monoxide rebreathing method determined by a meta-analysis. Br J Sports Med. 2013; 47: i31–9.
24. Gore CJ, Rodriguez FA, Truijens MJ, Townsend NE, Stray-Gundersen J, Levine BD. Increased serum erythropoietin but not red cell production afer 4 wk of intermittent hypobaric hypoxia (4,000–5,500 m). J Appl Physiol. 2006; 101: 1386–91. PMID: 16794028
25. Hopkins WG, Hawley JA, Burke LM. Design and analysis of research on sport performance enhancement. Med Sci Sports Exerc. 1999; 31: 472–85. PMID: 10188754
26. Hopkins W. How to interpret changes in an athletic performance test. Sportscience. 2004; 8: 1–7.

27. Hopkins WG, Schabort EJ, Hawley JA. Reliability of power in physical performance tests. Sports Med. 2001; 31: 211–34. PMID: 11286357

28. Bonetti DL, Hopkins WG. Variation in performance times of elite flat-water canoeists from race to race. Int J Sports Physiol Perf. 2010; 5: 210–7.

29. Pyne DB, Trewin C, Hopkins WG. Progression and variability of competitive performance of Olympic swimmers. J Sports Sci. 2004; 22: 613–20. PMID: 15370491

30. Smith TB, Hopkins WG. Variability and predictability of finals times of elite rowers. Med Sci Sports Exerc. 2011; 43: 2155–60. doi: 10.1249/MSS.0b013e31821d3f8e PMID: 21502896

31. Schmidt W, Prommer N. Effects of various training modalities on blood volume. Scand J Med.Sci. Sports 2008; 18: 57–69. doi: 10.1111/j.1600-0838.2008.00833.x PMID: 18665953

32. Lundby C, Calbert JA, Sander M, van Hall G, Mazzeo RS, Stray-Gundersen J et al. Exercise economy does not change after acclimatization to moderate to very high altitude. Scand J Med.Sci.Sports. 2007; 17: 281–91. PMID: 17501869

33. Conley DL, Krahenbuhl GS. Running economy and distance running performance of highly trained athletes. Med Sci Sports Exerc. 1980; 12: 357–60. PMID: 7453514

34. Daniels JT. A physiologist's view of running economy. Med Sci Sports Exerc. 1985; 17: 332–8. PMID: 3894870

35. Wilkinson M. Distinguishing between statistical significance and practica/clinical meaningfulness using statistical inference. Sports Med. 2014; 44.

36. Burton PR, Gurrin LC, Campbell MJ. Clinical significance not statistical significance: a simple Bayesian alternative to p values. J Epidemiol Comm Health. 1998; 52: 318–23.

37. Stang A, Poole C, Kuss O. The ongoing tyranny of statistical significance testing in biomedical research. Eur J Epidemiol. 2010; 25: 225–30. doi: 10.1007/s10654-010-9440-x PMID: 20339903

38. Stapleton C, S M.A., Atkinson G. The 'so what' factor: statistical versus clinical significance. Int J Sports Med. 2012; 30: 773–4.

39. Williams MN, Grajales GAG, Kurkiewiez D. Assumptions of multiple regression: correcting two misconceptions. Practical Assessment, Research and Evaluation. 2013; 18: 1–14.

40. Open_Science_Collaboration. Estimating the reproducibility of psychological science. Science. 2015; 349: 6251: aac4716.

41. Dienes Z. Using Bayes to get the most out of non-significant results. Front. Psych. 2014; 5: 781.

42. Gelman A, Stern H. The difference between "significant" and "not significant" is not itself statistically significant. Am Stat. 2006; 60: 328–31.

43. Verhagen J, Wagenmakers EJ. Bayesian tests to quantify the result of a replication attempt. J Exp Psych Gen. 2014; 143: 1457–75.