Articles in PresS. J Appl Physiol (February 5, 2015). doi:10.1152/japplphysiol.00714.2014

Individual response to exercise training – a statistical perspective

Anne Hecksteden,[1] Jochen Kraushaar,[1] Friederike Scharhag-Rosenberger,[1,2,3]

Daniel Theisen[4], Stephen Senn[5] and Tim Meyer[1]

[1] Institute of Sports- and Preventive Medicine, Saarland University, Saarbrücken, Germany

[2] Heidelberg University Hospital, National Center for Tumor Diseases (NCT), Heidelberg, Germany

[3] German University of Applied Sciences for Prevention and Health Management (DHfPG), Saarbrücken, Germany

[4] Sports Medicine Research Laboratory, Centre de Recherche Public de la Santé (CRP-Santé), Luxembourg, Luxembourg

[5] Competence Center in Methodology and Statistics, Centre de Recherche Public de la Santé (CRP-Santé), Luxembourg

Running head: Individual response – a statistical perspective

Corresponding author: Dr. Anne Hecksteden
Institute of Sports and Preventive Medicine
Saarland University
Campus Building B8.2
66123 Saarbrücken
Germany
Phone: 0049-681-302-3750
Fax: 0049-681-302-4296
eMail: a.hecksteden@mx.uni-saarland.de

Abstract

In the era of personalized medicine, interindividual differences in the magnitude of response to an exercise training program (subject-by-training interaction; “individual response”) have received increasing scientific interest. However, standard approaches for quantification and prediction remain to be established, probably due to the specific considerations associated with interactive effects, in particular on the individual level, as compared to the prevailing investigation of main effects. Regarding the quantification of subject-by-training interaction in terms of variance components, confounding sources of variability have to be considered. Clearly, measurement error limits the accuracy of response estimates and thereby contributes to variation. This problem is of particular importance for analyses on the individual level, because a low signal-to-noise ratio may not be compensated by increasing sample size (1 case). Moreover, within-subject variation in training efficacy may contribute to gross response variability. This largely unstudied source of variation may not be disclosed by comparison to a control group but calls for repeated interventions. A second critical point concerns the prediction of response. There is little doubt that exercise training response is influenced by a multitude of determinants. Moreover, indications of interaction between influencing factors of training efficacy lead to the hypothesis that optimal predictive accuracy may be attained using an interactive rather than additive approach. Taken together, aiming at conclusive inference and optimal predictive accuracy in the investigation of subject-by-training interaction entails specific requirements which are deducibly based on statistical principles but beset with many practical difficulties. Therefore, pragmatic alternatives are warranted.

Key words

variance components, prediction, determinant, interaction, moderator

Introduction

Everyday life experience overwhelmingly confirms that the magnitude of response to regular exercise training differs between individuals. This interindividual variability of training efficacy is reflected in the concepts of talent or trainability and translates to “subject-by-training interaction” in statistical terms. In fact, a high interindividual variability in the observed response to regular physical exercise is consistently reported by training studies, even within homogenous groups of previously untrained subjects and after fully compliant and supervised training. Moreover, this phenomenon is not only observed with regard to physical performance but also with virtually every physiological and health-related outcome measure (8).

This - so far unpredictable - individual training response presents a challenge on the practical as well as on the scientific level. Individual optimization of training prescriptions, statistical power in training studies and the interpretation of inconsistent results are some examples. The present manuscript aims to shed new light on problems of individual training response by discussing 3 main aspects. (A) The adequate assessment of individual response in experimental research. (B) The classification of individuals according to observed training effects (responder vs. non-responder). (C) Characteristics of predictive models for exercise training response. The general aim is to provide comprehensive insight into the specific methodological and statistical background of subject-by-training interaction, which differs in several important aspects from the paradigmatical analysis of main effects.

Individual response in experimental research

The vast majority of research questions addressed in training studies concerns the mean efficacy of interventions e.g. the mean difference in pre-post changes in VO_{2max} for the training group as compared to a control group. Consequently, the paradigmatic randomized controlled design (RCT) has been developed for the investigation of main effects. In the case of RCTs, the interindividual variability in training efficacy is a rather annoying characteristic which lowers standardised effect sizes and inflates the required sample size. However, in the

era of personalized medicine, individual training response has received increasing interest e.g. as the upper limit for genomic determination of training responsiveness (5) and a starting point to search for lifestyle-dependent moderators of training efficacy (25). However, it has to be kept in mind that methodological and statistical requirements for the valid assessment of interactive effects, including subject-by-training interaction, may differ from those for main effects.

Frequently, the high variability in training effects, which is consistently reported in training trials, is cited as evidence for differences in the responsiveness to exercise training between individuals (8, 21, 23). However, gross variability in pre-post changes does not represent subject-by-training interaction alone. Clearly, random (measurement) error in pre- and post-training assessments will contribute to observed differences in pre-post changes. This fact is generally acknowledged and taken into account today (20, 21, 23, 30). However, it is important to bear in mind that the inference from observed interindividual variability in training effects on subject-by-training interaction also relies on the (often implicit) assumption that the reproducibility of training effects on the individual level is high (32). Otherwise, the observed interindividual differences in training effects could just as well reflect an inconsistent intraindividual responsiveness to exercise. For example, it could be that some transient factor (e.g. amount of carbohydrate consumed during training bouts) (19) was influencing the effect of exercise for good or for ill. An overview of such moderators for the relationship between exercise training and training effects has been recently published by Mann et al. (25). Consequently, the response to exercise training may not only differ between people, but also within the same person on different occasions. A direct separation of consistent interindividual differences in training efficacy (subject-by-training interaction) from within-subject variability in training effects can only be achieved based on trials that repeatedly administer the same training intervention to the same participants (33). The underlying rationale for this requirement is similar to reliability trials which aim to quantify random within-subject variation in measured values and uncontestably require repeated measurements in the same subject (2, 21). However, from a practical point of view, the

implementation of training trials with repeated interventions interspersed by washout periods is, of course, beset by many difficulties (34). To the best of our knowledge, in the domain of sports medicine only one small trial with this design has been published so far almost 3 decades ago (35). To prepare for a detailed appraisal of alternative, more indirect approaches, a brief look at some relevant sources of variability is warranted.

Background: Relevant sources of variability

Random variation. Of course, measurement error and random (or yet unexplained) variability will affect pre- and post-training values for any outcome. According to statistical principles the contribution to the variance of an observed difference between measurements will be twice the within-subject variance for the individual measurement (21, 32), where this within-subject variance is the difference between observed and 'true' values. An Excel® spreadsheet is available online which demonstrates the influence of random variation by calculating the distribution of observed changes for the hypothetical case that all subjects are characterized by exactly the same true pre- and post-intervention values – and therefore experience the same true training effect. The background noise caused by random variation particularly impedes on the interpretation of single, observed pre- to post-training differences (e.g. the classification of individuals as non-responders or responders) (20) because a low signal-to-noise ratio may not be compensated for by increasing sample size (one individual). For a given number of subjects, relying on means from repeated tests before and after the training intervention (instead of a single measurement) reduces the random variation. In other words: Means of several observed values are a better estimate of the "true" value for an individual than a single observation. However, due to the additional cost and effort this option is seldom employed (6).

Between-subject variability. In a statistical sense between-subject variability is the difference in (true) values - not pre-post differences - between subjects. For example at the beginning of any training programme the true blood pressure (in the absence of treatment) will vary from individual to individual.

Subject-by-training interaction. This is what we actually mean when referring to "trainability", "talent" or "individual response": the consistent response to an exercise training program is dependent on the subject. It may also be described as the "between subject variability of true pre-post differences". It is important to bear in mind that genetic endowment will explain only part of an individual's training responsiveness (5, 8, 32). Stable lifestyle habits or epigenetic modifications are examples of other underlying mechanisms involved (14, 25).

Within-subject variability. That is the variability in the magnitude of pre-post differences when the same intervention is repeatedly administered to the same subject (after an adequate washout period). Importantly, within-subject variability refers to the variability of "true" pre-post differences devoid of random error in pre- and post- training measurements (see above). In analogy to individual measurements, this is the reliability of training effects. Consequently, repeated assessments are needed to quantify it, which in the case of pre-post differences obviously requires repetition of the intervention. From the perspective of logical rigor, the importance of considering within-subject variability in the quantification of individual response is beyond doubt and has been repeatedly brought forward for pharmacological interventions (33). However, for exercise training interventions little is known about the magnitude and practical relevance of this source of variation.

How to quantify subject-by-training interaction

Between-subject variability in (observed) training effects may be calculated in any training trial, and the standard deviation of pre-post changes provides a first indication of gross response variability (30). However, as outlined above, the standard deviation of training effects is an estimate of gross response variability which contains subject-by-training interaction as well as random error (in pre- and post- measurements) and within-subject variability in training response. An estimate for the contribution of random error may be derived from the standard deviation of a control group. By contrast, the separation of subject-by-training interaction and within-subject variability, which have both been introduced by the

158 training intervention, may not be achieved by a mere comparison to a control group. Instead,
159 separating reproducible from inconsistent components of training response requires the
160 repetition of the intervention. Just as the reliability of a single measurement may only be
161 assessed by repeating the measurement. In more technical terms, this translates to the
162 general rule that it is only possible to separate within-subject variation from interaction by
163 replication at the level at which it is desired to identify the interactive effect. For example, if
164 one wishes to identify sex-by-treatment interaction, a design in which members of both sexes
165 are given both treatments (for a two-treatment trial) is needed. On the other hand if subject-
166 by-treatment interaction is to be identified at the individual level, then individuals will have to
167 be repeatedly treated.

168 A simple simulation of within-subject variability. To exemplify the above
169 considerations, let us analyse 3 closely related hypothetical datasets. Table 1 presents the
170 changes in VO_{2max} (ΔVO_{2max}) for a "classical" randomized controlled training trial (RCT;
171 shaded part) as well as for two possible outcomes of a repeat intervention. In all cases, the
172 same numerical differences in VO_{2max} are observed in the respective groups. However, the
173 correlation between first (RCT) and repeat intervention is different. In version A of the repeat
174 intervention most values of ΔVO_{2max} are associated with the same person compared to the
175 first. This exemplifies high correlation between first and repeat intervention or, in other words,
176 low within-subject variability in training response. By contrast, in version B of the repeat
177 intervention, the individual values of VO2max are the same within each group, but have been
178 attributed differently to the different participants (most have been simply "shifted" to the next
179 subject). This exemplifies low correlation and high within-subject variability. Table 2 presents
180 the results of a simple linear mixed model applied to those datasets (dependent variable:
181 ΔVO_{2max}; fixed effect: group; random effect: subject-by-group interaction). The fixed effect
182 results demonstrate the same mean changes in VO_{2max} with a lower standard error for the
183 repeated trials. This is not surprising because the observed values of ΔVO_{2max} within each
184 group are the same in all 3 cases while the number of observations is higher in the repeated
185 trials. However, the estimates of subject-by-group interaction from the random effects are

dramatically different. In the case of the unrepeated “classical” RCT, most of the variability in ΔVO_{2max} is attributed to subject-by-group interaction. The same is true for version A of the repeated trial, where the individual magnitude of ΔVO_{2max} has been largely confirmed. By contrast, in version B of the repeated trial, the estimate for subject-by-group interaction is low and residual error (including within-subject variability in training effects here) accounts for most of the variability. Importantly, these scenarios are indiscernible after only the first training phase, i.e. without the repeated intervention. Of course, the example in this paragraph deliberately presents extreme cases of within-subject variability in training effects. However, we currently ignore the actual magnitude of this source of variability in the absence of a repeat intervention. This does not mean that individual response may not be estimated in data from unrepeated trials, but rather that the pertaining limitations need to be taken into account when interpreting quantitative estimates.

The direct approach: concurrent repetition of interventions. The straight-forward approach to implement the repetition of an intervention is within the respective trials themselves, of course after an adequate washout period (32, 33). Data analysis will be based on a linear mixed model approach with at least training fitted as fixed effect and subject identity as random effect. This approach also enables the calculation of confidence intervals for resulting variance components. However, the inclusion of repeated interventions in a training trial entails many difficulties which do not only concern cost and effort on the side of the research team, but also subject recruitment and compliance, as well as potential carry over and time effects (34). Therefore, in training trials with a primary research question focused on the main effect of the intervention, this design feature will be difficult to justify and implement on a routine basis.

The indirect approach: Separate reliability trials for training effects. A possible solution to this validity vs. feasibility dilemma is a separate “reliability trial of training effects”. This dedicated trial will be based on the repeated administration of the same training intervention with training phases separated by adequate washout periods. The resulting

quantitative estimate of within-subject variability in training response may then be used to arithmetically isolate subject-by-training interaction in data from subsequent studies with unrepeated (and therefore less extensive) designs. The most evident advantage of this approach would be efficiency, resulting from gaining the information only once and then using it further on. Another closely linked potential advantage may be the precision in the estimate of within-subject variance. If the effort of repetition has to be undertaken just once, a larger number of subjects can be included, leading to a more precise estimate. This is a very relevant advantage in favour of external repetition considering the small number of participants in many training trials (and the high dependence of precision in the estimates of variance components on the number of degrees of freedom). On the other hand, the most imminent disadvantage seems to be generalizability - an issue which is aggravated by the diversity of exercise interventions and outcome measures. Also, various sorts of bias may come into play (e.g. standardisation measures taken, subject characteristics, time and assay effect). To date it seems difficult to estimate the relevance of these concerns for exercise training interventions.

Going without repeated interventions – how far will we get? Some authors suggest quantifying individual response by subtracting the variance of pre-post differences in a control group from that in the experimental group (21, 23). In fact, this approach elegantly separates variation due to random error (which is present in either group) from variation due to the intervention (which is present in the training group only). Moreover, if the analysis is based on a general mixed model, confidence intervals for the imputed variance components may be specified (21, 23). However, as demonstrated above, it has to be kept in mind that the remaining variability in the experimental group still represents subject-by-training interaction as well as within-subject variability in training response which have both been introduced by the training intervention.

Repeated testing during the training phase as a surrogate for repeated interventions. A sensible amendment to the approach just discussed may be the repeated testing of

outcome measures during the course of a single uninterrupted intervention period (Figure 1). Beyond reducing the impact of measurement error, this provides repeated pre-post differences for shorter (segmental) intervention periods. When fit by a linear mixed model with random intercept and slopes (12, 36, 40), subject-by-training interaction (that is between-subject variability in slopes), may be separated from within-subject variability (in segmental slopes) (31). This approach approximates the validity of truly repeated interventions for the assessment of individual response while still avoiding major aspects of their complexity e.g. washout period and prolonged trial duration. However, some limitations have to be accepted which are mostly caused by the accumulation of tests within the limited duration of a typical training study (13). Particularly, a lack in temporal separation of measurements may lead to high degrees of autocorrelation and violate the assumption of random errors. Moreover, training adaptation over the whole training period may not always be linear (29) and baselines for segmental training periods will differ because a part of the adaptation has already occurred. These problems may be addressed by including the segmental baseline as a covariate and by choosing a non-linear model if appropriate. From a physiological perspective, frequent exercise tests may interfere with the training intervention. Lastly, from the perspective of study conduct, repeated testing will lead to a marked increase in cost and effort.

As an intermediate conclusion, when theoretical and pragmatical arguments are taken together, there is probably no universal solution to the quantification of individual response for every training trial. However, it is essential to specify the degree of detail in the separation of sources of variability which was achieved with the methodological and statistical approach employed in a particular trial. This may range from gross variability in pre-post differences for an uncontrolled trial, to the formal quantification of subject-by-training interaction when estimates of random error as well as within-subject variability are available. An overview is provided in Table 3.

Responders and Non-Responders: Classification of individuals

A classical way to display the variation in pre-post differences in a training trial consists in plotting the magnitude of training effects (change from pre- to post-intervention values in an outcome measure) against the cumulative number of observations (Figure 2 – with permission). This approach has been introduced by Bouchard et al. using data from the HERITAGE family study (6, 8). A striking feature of their figures is that, despite a significant mean training effect, the distribution not only approaches but also crosses the x-axis (zero change) for all parameters examined. This lack of (or even adverse) change after an intervention has been termed “non-response”.

In the meantime, “non-response” or even “adverse response” to regular physical exercise has been addressed by several authors (6, 7, 10, 30, 37, 39). A main challenge in this field is the accurate and adequate classification of individuals according to the magnitude of training response. The sources of variation discussed above lead to an inevitable uncertainty regarding the difference between the observed and the true training effect for the individual subject. Moreover, the rationale for meaningful threshold values between responders and non-responders is still a matter of debate. In particular, there is no consensus on whether to define “response” by the (probable) presence of clinically relevant change (17, 22) or of clearly measurable change (7, 30). Consequently, the definitions of non-response or non-responders differ considerably.

0 difference as threshold value: A straight-forward approach is to define non-response as a difference from pre- to post-training values of 0 or less (for outcomes that usually increase with training). Although they do not give a formal definition, this seems to be the concept that Bouchard et al. had in mind in their early work (6). However, this intuitive approach has a serious shortcoming: it does not take the limited accuracy of observed training effects into account. Even if all subjects experienced the same, true training effect, there would still be random variation in the observed changes due to random variability. The proportion of subjects with an observed training effect of zero or less caused by random variability of individual measurements alone may be calculated from the within-subject

variability of the outcome measure (reliability) and the true training effect (mean). This circumstance is illustrated by the Excel® spreadsheet in supplemental digital content 1. Bouchard et al. were obviously aware of this problem and emphasize their efforts to optimize the reliability of measurements.

Alternative threshold values: Another point of criticism with regard to the aforementioned definition of non-response has been the fixed cut-off value (0). Instead, the borderline between trivial and substantial effects has been suggested as a preferable criterion for the discrimination of responders and non-responders (17, 30). From the perspective of practical application, this cut-off is represented by the minimum "clinically relevant change" (17) or "smallest worthwhile difference" (22). From a statistical point of view, the coefficient of variation (CV) for the respective outcome (and methodology) delineates differences that may be expected as a result of random variation from (probably) true changes (21, 23, 30). Both approaches raise the bar for training effects and make the classification of subjects as responders more conservative. However, the above shortcoming, that is the uncertainty in the classification of an individual due to random variation in observed values and within-subject response variability, will apply regardless of the chosen cut-off value.

Estimate of uncertainty in the classification: As outlined above, the accuracy of observed training effects is inevitably limited – even if all methodological devices for the reduction and control of random error are employed (e.g. standardisation, averaging of multiple measurements, control group). However, the ensuing uncertainty in the classification of an individual as responder or non-responder may be quantified by appropriate statistical techniques. In particular, linear mixed modelling enables the specification of confidence limits for each individual's response. The combination of this approach with prefixed limits of practical relevance as suggested by Hopkins et al. (20, 21) is the most informative option for the classification of responders and non-responders based on a single observed training effect. However, it has to be kept in mind that the underlying assumption of negligible within-

subject variability in training efficacy is untested so far. A detailed introduction into principles (12, 36, 40) and specific application (20, 21) of linear mixed modelling, is beyond the scope of this manuscript. Interested readers are referred to the references provided above.

Quantiles. An alternative approach to the definition of non-response is based on quantiles (37, 39). For example, the quarter of subjects with the lowest training effect may be regarded as non-responders (and the quarter with the highest training effects as high-responders). Quantiles offer a convenient way to contrast balanced subgroups with marked differences in training effects. However, because a predefined percentage of subjects will be regarded as non-responders, it offers little insight into the distribution and variability of training effects.

Allowing for several outcome measures. In the above paragraphs individual training response has been discussed in general, without specifying the criterion outcome(s). In studies which specifically addressed the issue, VO_{2max} is the most common outcome measure (6, 8, 30, 39). However, similar variability has been found “wherever it was looked for” (8). In response to the uncertainty associated with the interpretation of an observed individual difference (see above), it has been suggested to consider several outcome measures with similar meaning (30). The authors of this study used the coefficient of variation (CV) as cut-off value for the discrimination of non-responders and report differences in the rate of non-response for several parameters of physical performance. However, each of their 18 subjects responded in at least one parameter. The consideration of several outcomes adds a qualitative dimension to the assessment of individual response. However, it does not account for the uncertainty in the classification for individual parameters. Moreover, the multiplication of parameters is associated with over-reporting of partial non-response and underreporting of complete non-response.

Training response – quantification and causality. A shared characteristic of the above definitions is the classification of subjects according to the magnitude of their individual pre-post difference. Of course, comparing the phenotype of an individual before and after an

intervention is the one-and-only way to evaluate his/her response. However, this exclusive focus on intraindividual differences and the ensuing lack of a head-to-head comparison to a control group entails important limitations for the interpretation of “response”. To see this, consider the two distributions in Figure 3. For both the experimental and the control group it is clear that many have a difference from baseline that is negative, even though the number is larger for the control group. However, the experimental distribution is the same as the control group distribution but shifted to the right. Thus every member of the experimental group, whether the difference from baseline is negative or not, can be matched with a corresponding member of the control group. Compared to this corresponding value from the control group each member of the experimental group has had an identical benefit. As a conclusion, one should be very cautious about giving “response” used in this way a causal interpretation, because, as in every uncontrolled approach, we do not know what would have happened without the intervention.

Training response – universal or protocol specific. Beyond definitions and study design, the classification of individuals as responders or non-responders is also dependent on the characteristics of training intervention and criterion outcome. Importantly, the timeline of training adaptations differs considerably depending on the outcome measure in question e.g. submaximal measures of physical performance seem to respond faster than VO_{2max} (29), left ventricular wall thickness faster than left ventricular diameter (1) and resting blood pressure faster than other cardiovascular risk factors (26). Therefore, an individual who seems to be a non-responder after a typical training phase of 3-6 months (13) may well have responded to a longer intervention or for an outcome with a faster time course. Moreover, there is evidence that the proportion of non-responders can vary depending on the intervention type. For example, the rate of non-responders seems to be lower for high intensity interval training than for continuous exercise of moderate intensity (4). In consideration of these aspects, the classification of study participants as responders or non-responders has to be acknowledged as being time and protocol specific.

Predicting (and optimizing) training response

In the above paragraphs individual response has been discussed from an analytical, post-hoc perspective, aiming at the identification of variance components. This perspective is of fundamental importance for the conduct and interpretation of related research. Moreover, between- and within-subject variance components delineate the possible impact of subject-inherent and modifiable determinants. However, with regard to practical application, the more important perspective concerns the prediction and, ultimately, individual optimization of training effects. To this end, relevant physiological determinants of training effects need to be known. So far, research in this direction has largely focused on genetic factors which seem to account for roughly 50% of the interindividual variability in training effects (5, 9, 11, 27, 28, 38) (Figure 4a). This plausible primary research focus has also been pursued in other cases of marked interindividual differences in the effects of an intervention e.g. salt restriction as a means to lower blood pressure (24). However, like most phenotypical changes, training effects result from a larger set of determinants comprising subject inherent, exercise associated as well as external factors (e.g. nutrient intake around training; Figure 4b). A comprehensive overview of such moderators of training response has been recently published (25). A prominent example for a strong external moderator of treatment efficacy from the field of pharmacology is the interaction of grapefruit juice with a multitude of medications (15). However, a list of explanatory variables is only a first step towards the prediction of training effects. Specification of a quantitative predictive model also requires knowing their independent impact as well as potential interactions within the set of determinants. In other words, the concept of training response underlying the predictive model (univariable vs. multivariable; additive vs. interactive; cp. Figure 4) has to be specified.

Univariable or multivariable? As already stated above, there may be little doubt that training efficacy is influenced by a multitude of moderators (25). This multivariable character is the foundation of sensible standardization measures, randomization and stratification. Therefore, the paradigmatic randomized controlled training trial is based on a multivariable concept of training response, even if its statistical analysis is frequently not.

Additive or interactive? By contrast to numerous studies on individual moderators of training efficacy, interdependencies within the set of determinants have rarely been addressed. However, in some cases, an interaction is obvious. A simple example is the interaction between training mode and carbohydrate availability around the time of training. While carbohydrate supply is known to support the effect of strength training (19), it probably impairs the adaptive response to endurance exercise (3, 18). Moreover, a complete independence seems particularly unlikely for exercise training (as compared to pharmacological interventions) because exercise has to generate effects “per vias naturales”, that is via the intricate homeostatic signal transduction network of the cell, without taking pharmaceutical shortcuts. In fact, if the magnitude (or even direction) of influence for one determinant depends on the values of others, such interactions may lead to several highly relevant consequences, the most pertinent one being meaningful factor combinations. The aforementioned effect of carbohydrate intake during strength- or endurance training is a very simple example of such a “pattern”. It is noteworthy that if the various determinants are studied separately, it is not possible to analyze their interaction. Importantly, this principal restriction also applies to metaanalytical approaches. In other words, if the determinants of training response do interact, a metaanalytical integration of studies on individual influencing factors must be interpreted very cautiously.

Taken together there may be little doubt, that training response is influenced by a multitude of determinants and that, at least in some cases, the influence of one determinant depends on the value of others. These characteristics are represented by a multivariable, interactive concept of training response (as depicted in Figure 4c). However, generating the dataset for the deduction of a respective predictive model requires a large-scale, cross-over training study, which permits simultaneous assessment of a comprehensive set of moderators. These characteristics approach the limit of the factorial design type (16). To the best of our knowledge, no attempt has been published in this direction. Therefore, the most accurate predictive approach for the near future will probably be based on an individual’s

genetically determined responsiveness to exercise and the main effect of major moderators (Figure 4b).

Conclusion

The quantification and prediction of individual response to exercise training is associated with specific considerations and challenges as compared to the investigation of main effects. Although, from the perspective of logical rigor, the resulting methodological and statistical requirements are beyond doubt, in most cases it will not be possible to implement them to their full extent. This concerns in particular the need for repeated interventions to separate subject-by-training interaction from within-subject variability in training effects. Consequently, pragmatic alternatives are warranted, depending on research question and available data. However, the achieved degree of detail in the separation of sources of variability has to be carefully specified as well as potential assumptions and limitations. An overview is provided in Table 3.

With regard to the prediction of training response it seems clear that a multivariable, interactive model most closely reflects physiological circumstances. However, the deduction of such a model requires the assessment of all relevant explanatory variables in a single experimental approach. This could only be realized in a large scale, cross-over training study approaching the complexity of the factorial design type. To date predictive models can only be based on the main effect of relevant determinants. However, research into the interdependence of determinants is clearly warranted.

Grants

Stephen Senn's work was funded by EU FP7 2013 grant number 602552 as part of the IDEAL project.

Disclosures

The authors declare that there is no true or apparent conflict of interest.

References

1. **Arbab-Zadeh A, Perhonen M, Howden E, Peshock RM, Zhang R, Adams-Huet B, Haykowsky MJ, and Levine BD**. Cardiac Remodeling in Response to 1 Year of Intensive Endurance Training. *Circulation* 2014.
2. **Atkinson G, and Nevill AM**. Statistical methods for assessing measurement error (reliability) in variables relevant to sports medicine. *Sports Med* 26: 217-238, 1998.
3. **Baar K**. Nutrition and the adaptation to endurance training. *Sports Med* 44 Suppl 1: S5-12, 2014.
4. **Bacon AP, Carter RE, Ogle EA, and Joyner MJ**. VO2max trainability and high intensity interval training in humans: a meta-analysis. *PLoS One* 8: e73182, 2013.
5. **Bouchard C**. Genomic predictors of trainability. *Exp Physiol* 97: 347-352, 2012.
6. **Bouchard C, An P, Rice T, Skinner JS, Wilmore JH, Gagnon J, Perusse L, Leon AS, and Rao DC**. Familial aggregation of VO(2max) response to exercise training: results from the HERITAGE Family Study. *J Appl Physiol* 87: 1003-1008, 1999.
7. **Bouchard C, Blair SN, Church TS, Earnest CP, Hagberg JM, Hakkinen K, Jenkins NT, Karavirta L, Kraus WE, Leon AS, Rao DC, Sarzynski MA, Skinner JS, Slentz CA, and Rankinen T**. Adverse metabolic response to regular exercise: is it a rare or common occurrence? *PLoS One* 7: e37887, 2012.
8. **Bouchard C, and Rankinen T**. Individual differences in response to regular physical activity. *Med Sci Sports Exerc* 33: S446-451; discussion S452-443, 2001.
9. **Bouchard C, Sarzynski MA, Rice TK, Kraus WE, Church TS, Sung YJ, Rao DC, and Rankinen T**. Genomic predictors of the maximal O uptake response to standardized exercise training programs. *J Appl Physiol* 110: 1160-1170, 2011.
10. **Boule NG, Weisnagel SJ, Lakka TA, Tremblay A, Bergman RN, Rankinen T, Leon AS, Skinner JS, Wilmore JH, Rao DC, and Bouchard C**. Effects of exercise training on glucose homeostasis: the HERITAGE Family Study. *Diabetes Care* 28: 108-114, 2005.

11. **Bray MS, Hagberg JM, Perusse L, Rankinen T, Roth SM, Wolfarth B, and Bouchard C**. The human gene map for performance and health-related fitness phenotypes: the 2006-2007 update. *Med Sci Sports Exerc* 41: 35-73, 2009.

12. **Burton P, Gurrin L, and Sly P**. Extending the simple linear regression model to account for correlated responses: an introduction to generalized estimating equations and multi-level mixed modelling. *Stat Med* 17: 1261-1291, 1998.

13. **Cornelissen VA, and Fagard RH**. Effects of endurance training on blood pressure, blood pressure-regulating mechanisms, and cardiovascular risk factors. *Hypertension* 46: 667-675, 2005.

14. **Ehlert T, Simon P, and Moser DA**. Epigenetics in sports. *Sports Med* 43: 93-110, 2013.

15. **Farkas D, and Greenblatt DJ**. Influence of fruit juices on drug disposition: discrepancies between in vitro and clinical studies. *Expert Opin Drug Metab Toxicol* 4: 381-393, 2008.

16. **Fisher R**. The Arrangement of Field Experiments. *Journal of the Minitry of Agriculture of Great Britain* 33: 503-513, 1926.

17. **Guyatt GH, Juniper EF, Walter SD, Griffith LE, and Goldstein RS**. Interpreting treatment effects in randomised trials. *BMJ* 316: 690-693, 1998.

18. **Hawley JA, and Burke LM**. Carbohydrate availability and training adaptation: effects on cell metabolism. *Exerc Sport Sci Rev* 38: 152-160, 2010.

19. **Hawley JA, Burke LM, Phillips SM, and Spriet LL**. Nutritional modulation of training-induced skeletal muscle adaptations. *J Appl Physiol* 110: 834-845, 2011.

20. **Hopkins WG**. How to Interpret Changes in an Athletic Performance Test. *sportsciorg/jour/04/wghtestshtm* acessed: 14.06.2013

21. **Hopkins WG**. Measures of reliability in sports medicine and science. *Sports Med* 30: 1-15, 2000.

22. **Hopkins WG, Hawley JA, and Burke LM**. Design and analysis of research on sport performance enhancement. *Med Sci Sports Exerc* 31: 472-485, 1999.

510 23. **Hopkins WG, Marshall SW, Batterham AM, and Hanin J**. Progressive statistics for
511 studies in sports medicine and exercise science. *Med Sci Sports Exerc* 41: 3-13,
512 2009.
513 24. **Kelly TN, and He J**. Genomic epidemiology of blood pressure salt sensitivity. *J*
514 *Hypertens* 30: 861-873, 2012.
515 25. **Mann TN, Lamberts RP, and Lambert MI**. High Responders and Low Responders:
516 Factors Associated with Individual Variation in Response to Standardized Training.
517 *Sports Med* 2014.
518 26. **Murray A, Delaney T, and Bell C**. Rapid onset and offset of circulatory adaptations
519 to exercise training in men. *J Hum Hypertens* 20: 193-200, 2006.
520 27. **Rankinen T, Roth SM, Bray MS, Loos R, Perusse L, Wolfarth B, Hagberg JM,**
521 **and Bouchard C**. Advances in exercise, fitness, and performance genomics. *Med*
522 *Sci Sports Exerc* 42: 835-846, 2010.
523 28. **Roth SM, Rankinen T, Hagberg JM, Loos RJ, Perusse L, Sarzynski MA, Wolfarth**
524 **B, and Bouchard C**. Advances in exercise, fitness, and performance genomics in
525 2011. *Med Sci Sports Exerc* 44: 809-817, 2012.
526 29. **Scharhag-Rosenberger F, Meyer T, Walitzek S, and Kindermann W**. Time course
527 of changes in endurance capacity: a 1-yr training study. *Med Sci Sports Exerc* 41:
528 1130-1137, 2009.
529 30. **Scharhag-Rosenberger F, Walitzek S, Kindermann W, and Meyer T**. Differences
530 in adaptations to 1 year of aerobic endurance training: individual patterns of
531 nonresponse. *Scand J Med Sci Sports* 22: 113-118, 2012.
532 31. **Schielzeth H, and Forstmeier W**. Conclusions beyond support: overconfident
533 estimates in mixed models. *Behav Ecol* 20: 416-420, 2009.
534 32. **Senn S**. Individual response to treatment: is it a valid assumption? *BMJ* 329: 966-
535 968, 2004.

33. **Senn S, Rolfe K, and Julious SA**. Investigating variability in patient response to treatment--a case study from a replicate cross-over study. *Stat Methods Med Res* 20: 657-666, 2011.

34. **Senn SJ**. Cross-over Trials in Clinical Research. *John Wiley & Sons, Chichester, United Kingdom* 2. edition: 2002.

35. **Simoneau JA, Lortie G, Boulay MR, Marcotte M, Thibault MC, and Bouchard C**. Effects of two high-intensity intermittent training programs interspaced by detraining on human skeletal muscle and performance. *Eur J Appl Physiol Occup Physiol* 56: 516-521, 1987.

36. **Sullivan LM, Dukes KA, and Losina E**. Tutorial in biostatistics. An introduction to hierarchical linear modelling. *Stat Med* 18: 855-888, 1999.

37. **Timmons JA, Jansson E, Fischer H, Gustafsson T, Greenhaff PL, Ridden J, Rachman J, and Sundberg CJ**. Modulation of extracellular matrix genes reflects the magnitude of physiological adaptation to aerobic exercise training in humans. *BMC Biol* 3: 19, 2005.

38. **Timmons JA, Knudsen S, Rankinen T, Koch LG, Sarzynski M, Jensen T, Keller P, Scheele C, Vollaard NB, Nielsen S, Akerstrom T, MacDougald OA, Jansson E, Greenhaff PL, Tarnopolsky MA, van Loon LJ, Pedersen BK, Sundberg CJ, Wahlestedt C, Britton SL, and Bouchard C**. Using molecular classification to predict gains in maximal aerobic capacity following endurance exercise training in humans. *J Appl Physiol* 108: 1487-1496, 2010.

39. **Vollaard NB, Constantin-Teodosiu D, Fredriksson K, Rooyackers O, Jansson E, Greenhaff PL, Timmons JA, and Sundberg CJ**. Systematic analysis of adaptations in aerobic capacity and submaximal energy metabolism provides a unique insight into determinants of human aerobic performance. *J Appl Physiol* 106: 1479-1486, 2009.

40. **Zuur AF, Ieno, Elena N., Smith, Graham M.** Introduction to mixed modelling Chapter 8 in. *Analysing Ecological Data* Springer: 125-142, 2007.

565 **Tables**

566

Table 1. Hypothetical trials - Raw data

	RCT			Repetition	
Name	**Subject ID**	**Group**	**ΔVO_{2max} RCT**	**ΔVO_{2max} Repeat A**	**ΔVO_{2max} Repeat B**
Anne	**1**	**0**	**-1.81**	-1.81	1.25
Berta	**2**	**0**	**5.25**	5.25	-1.81
Clara	**3**	**0**	**2.59**	2.59	5.25
Dieter	**4**	**0**	**8.40**	8.40	2.59
Ernst	**5**	**0**	**5.96**	5.96	8.40
Fabienne	**6**	**0**	**1.47**	1.47	5.96
Gustav	**7**	**0**	**10.53**	10.53	1.47
Heike	**8**	**0**	**1.25**	1.25	10.53
Isabell	**9**	**0**	**5.31**	4.15	5.31
Jana	**10**	**0**	**4.15**	5.31	4.15
Karl	**11**	**1**	**15.89**	15.89	0.02
Lena	**12**	**1**	**28.42**	28.42	15.89
Maria	**13**	**1**	**8.08**	8.08	28.42
Niko	**14**	**1**	**19.32**	19.32	8.08
Olga	**15**	**1**	**6.55**	6.55	19.32
Peter	**16**	**1**	**6.28**	6.28	6.55
Robert	**17**	**1**	**8.70**	8.70	6.28
Sarah	**18**	**1**	**0.02**	0.02	8.70
Tina	**19**	**1**	**14.56**	21.42	14.56
Uwe	**20**	**1**	**21.42**	14.56	21.42

Group codes: 0: control; 1: training;
Shaded part: standard randomized controlled trial (RCT)
ΔVO_{2max}: Pre-Post difference in maximum oxygen uptake
Repeat A: Hypothetical second (repeated) intervention phase with low within-subject variability in training effects (ΔVO_{2max} is identical to the first phase (RCT) for all but the last 2 subjects of each group which are interchanged.)
Repeat B: Hypothetical second (repeated) intervention phase with high within-subject variability in training effects (ΔVO_{2max} is shifted downwards by one person compared to the first phase (RCT) except for the last 2 subjects of each group.)

567

568

569 **Table 2**

570

Table 2. Hypothetical trials – Results			
	Unrepeated Trial	**Repeated Trial**	
		A	B
Fixed Effect Δ VO_{2max} ($ml*kg^{-1}*min^{-1}$); means ± standard error			
Group 1	4.3 ± 1.1	4.3 ± 0.8	4.3 ± 0.8
Group 2	12.9 ± 2.7	12.9 ± 1.9	12.9 ± 1.9
Random Effects Standard deviations ($ml*kg^{-1}*min^{-1}$) ± standard error			
ID*Group	6.4 #	6.4 ± 3.7	1.9 ± 3.1
Error	1.0 #	1.5 ± 0.9	6.1 ± 3.4
Group 1: controls; Group 2: training Δ VO_{2max}: Pre-post difference in maximum oxygen uptake Mixed linear model: Fixed effect: group (training vs. control) Random effect: subject-by-group interaction (ID*group) #: Standard errors will not be reported due to the redundancy in the model			

571

572 **Table 3**

Table 3. Assessment of subject-by-training interaction with different trial designs

Design type	Statistical approach	Separation of subject-by-training interaction	Specific limitations
Uncontrolled			
Standard	Descriptive (30)	No Only gross variability (effect range and variation)	Inability to establish causality
with repeated testing	Mixed linear model (random slopes) (31)	Yes	Possible concomitants of test accumulation (autocorrelation; for exercise testing: interference with training intervention)
Controlled			
Standard RCT	Difference in variation between training and control group (via mixed linear model) (21, 23)	Partially Separates training associated variability (incl. subject-by-training interaction) from ubiquitous sources of variation (e.g. measurement error)	Difference in variation between the training and control groups is neither necessary nor sufficient for subject-by-training interaction to be present. Untested assumption of low within-subject variability in training effects.
with repeated testing	Mixed linear model (random slopes) (31)	Yes	Possible concomitants of test accumulation (autocorrelation; interference with training)
with repeated intervention	Mixed linear model (33)	Yes	Cost and effort; approx. 3-fold total trial duration
Arithmetical integration of variance components from previous work Quantitative estimates of specific sources of variation (e.g. measurement error or within-subject variability in training effects) may be utilized during data analysis to overcome limitations imposed by the respective trial design. However, provided that such estimates are available at all, their applicability to the specific population, outcome, intervention and protocol has to be critically appraised.			
Mixed linear modelling Mixed linear models are based on the principles of linear regression. However, main (fixed) effects and components of variability (random effects) are modelled separately. Due to the possibility to include subject identity as a random effect, mixed models are particularly suitable for the analysis of longitudinal studies and the analysis of between- and within-subject variability. A detailed introduction into principles (12, 36, 40) and specific application (20, 21) of linear mixed modelling, is beyond the scope of this manuscript.			

Figure Captions

Fig. 1 Individual courses of changes in maximum oxygen uptake

a Pre- and Post-intervention assessment only

b With additional assessments during the intervention period

Data pertain to subjects 15-20 from table 1.

Fig. 2 Individual differences in increase in VO_{2max} with training (from (6) with permission)

Fig. 3 Distribution of pre-post differences in a hypothetical randomized controlled training trial with equal variation in the training and control groups, respectively.

Fig. 4 Concepts of training response. Possible schematic representations depending on their taking into account multiple determinants and their interaction.

a Nature-and-Nurture- The parsimonious concept

b Multivariable concept without interaction between determinants

c Multivariable concept allowing for interactions between determinants

Figure 1

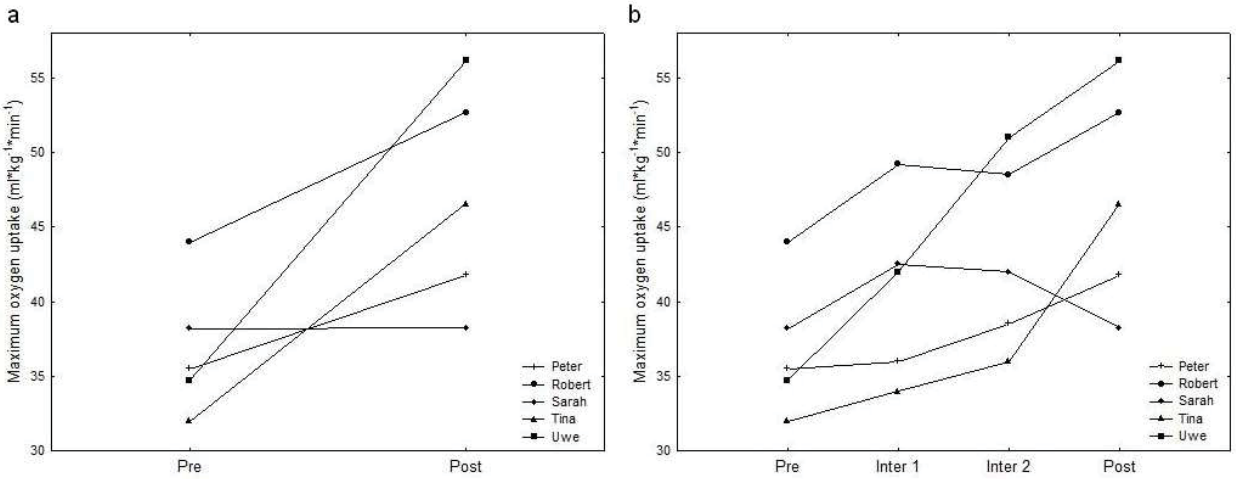

a
b
Maximum oxygen uptake (ml*kg-1*min-1)
55
50
45
40
35
30
Pre
Post
Inter 1
Inter 2
Peter
Robert
Sarah
Tina
Uwe

Figure 2

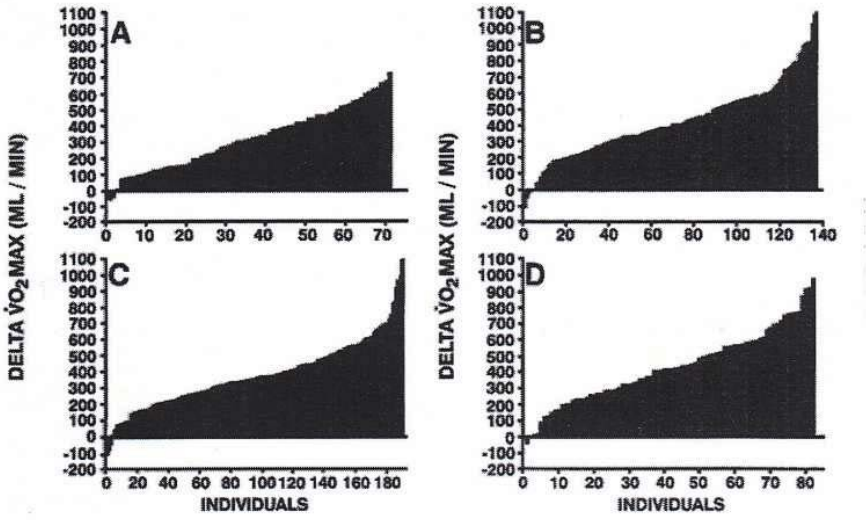

A
B
C
D
DELTA $\dot{V}O_2$MAX (ML / MIN)
DELTA $\dot{V}O_2$MAX (ML / MIN)
INDIVIDUALS
INDIVIDUALS

Figure 3

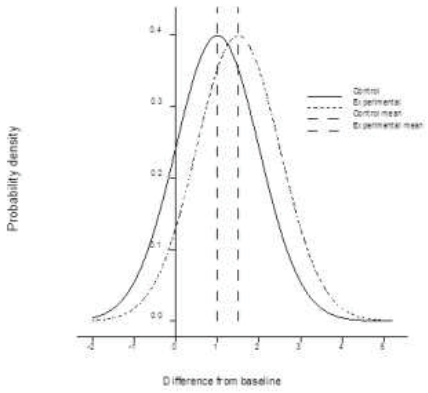
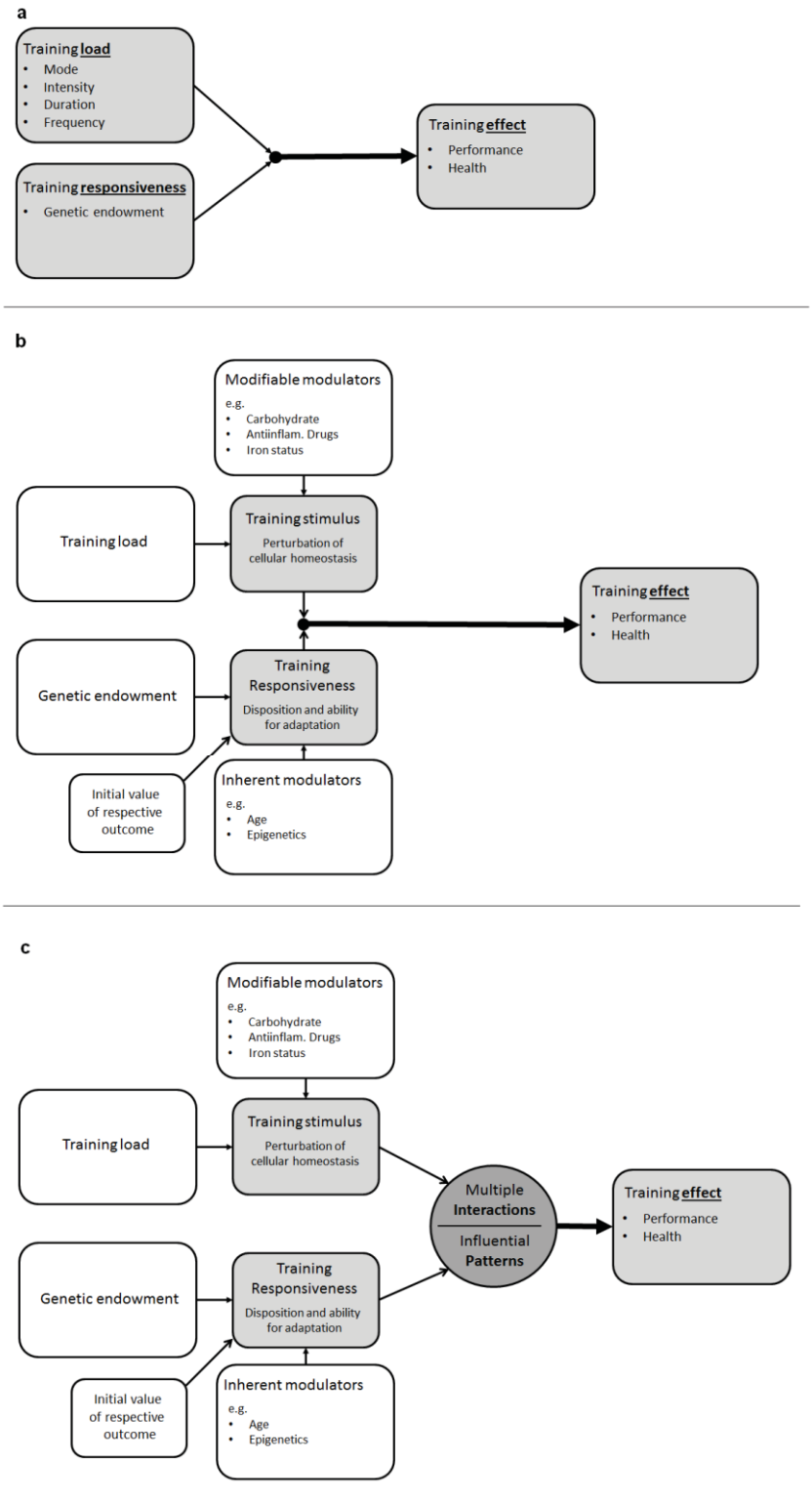

Probability density
Difference from baseline
Control
Experimental
Control mean
Experimental mean

Figure 4


a
Training load
Mode
Intensity
Duration
Frequency
Training responsiveness
Genetic endowment
Training effect
Performance
Health
b
Modifiable modulators
e.g.
Carbohydrate
Antiinflam. Drugs
Iron status
Training load
Training stimulus
Perturbation of cellular homeostasis
Training effect
Performance
Health
Genetic endowment
Training Responsiveness
Disposition and ability for adaptation
Initial value of respective outcome
Inherent modulators
e.g.
Age
Epigenetics
c
Modifiable modulators
e.g.
Carbohydrate
Antiinflam. Drugs
Iron status
Training load
Training stimulus
Perturbation of cellular homeostasis
Multiple Interactions
Influential Patterns
Training effect
Performance
Health
Genetic endowment
Training Responsiveness
Disposition and ability for adaptation
Initial value of respective outcome
Inherent modulators
e.g.
Age
Epigenetics