Improving Meta-analyses in Sport and Exercise Science

Will G Hopkins

Institute of Sport Exercise & Active Living, Victoria University, Melbourne, Australia. Email.
Reviewer: Alan M Batterham, School of Health and Social Care, University of Teesside, Middlesbrough, UK.

Flaws in recent meta-analyses in our disciplines include extraction of inappropriate measures or units from published studies, inadequate documentation of the standard errors used to weight the effects, deficiencies in accounting for heterogeneity in the effect between studies, inappropriate treatment of multiple measures from the same studies, absent or deficient analysis of publication bias, and failure to interpret the clinical or practical importance of the meta-analyzed mean and heterogeneity. More attention to these flaws by authors, reviewers and editors will raise the quality of meta-analyses accepted for publication. Until then researchers and practitioners should use published meta-analyses with caution, especially those lacking adequate analysis of heterogeneity: use such analyses primarily as convenient reference lists to find studies from which to draw conclusions about the effect in settings similar to your own. KEYWORDS: magnitude-based inference, mixed models, random-effect meta-regression, systematic reviews.

Reprint pdf · Reprint docx

See also the article/slideshow on meta-analysis. The article/slideshow on the replication crisis also summarizes meta-analysis.

Introduction

As a reviewer of meta-analyses for several journals, I have noticed a worrying trend: authors of submitted manuscripts often justify their flawed analyses by citing published articles with similar flaws. My aim in this article is to identify flaws in several recent meta-analyses in the hope of raising the quality of submitted and published meta-analyses in our disciplines. Numerous flaws in a recent meta-analysis by Coquart et al. (2016a) provided the stimulus for writing the article. I performed a search for meta-analyses published since 2011 over a wide range of topics to find more examples of errors and omissions (Burden et al., 2015; Chu et al., 2016; Hume et al., 2015; Josefsson et al., 2014; Peterson et al., 2011; Salvador et al., 2016; Soomro et al., 2015;

Tomlinson et al., 2015; Wright et al., 2015). One of the meta-analyses (Soomro et al., 2015) turned out to have only one minor flaw. This article represents a summary of the flaws and advice on how to avoid them. I submitted the article to Sports Medicine in 2016, but the editor decided that he would prefer a meta-analysis of meta-analyses, something I did not have time or inclination to write. Instead, I wrote a letter to the editor about the Coquart et al. study (Hopkins, 2016), to which the authors responded by showing that they did not understand the difference between confidence limits and limits of agreement, by continuing to cite flawed meta-analyses as justification, and by claiming incorrectly that heterogeneity statistics are not always

needed (Coquart et al., 2016b).

After setting aside this article for a year, I decided to submit it for publication in Sportscience. I have identified the flaws under subheadings that reflect the temporal order in which a meta-analysis is usually analyzed. My assertions about the wrong and right ways to do a meta-analysis are based on an article/slideshow I published in 2004, which I have updated regularly (Hopkins, 2004), and on a more generic guidelines article (Hopkins et al., 2009). The right ways are exemplified (I hope) in recent meta-analyses I have co-authored (Bonetti and Hopkins, 2009; Braakhuis and Hopkins, 2015; Carr et al., 2011; Cassar et al., 2017; McGrath et al., 2015; Snowling and Hopkins, 2006; Somerville et al., 2016; Vandenbogaerde and Hopkins, 2011; Weston et al., 2014). The Cochrane handbook (Higgins and Green, 2011) is also a useful source of wisdom and software for simple meta-analyses, but only a powerful mixed modeling procedure (e.g., Proc Mixed in the Statistical Analysis System) can provide the appropriate random effects to deal with repeated measurement represented by multiple study-estimates from the same or different subjects in each study.

Expressing Effects in the Same Metric

A meta-analyzed effect is a weighted mean of the effect across all selected studies. All effects must therefore be expressed as the same kind of measure in the same units for the mean to be meaningful. Most of the meta-analyses I chose for this article used unsuitable approaches here.

For differences or changes in means, a common approach is to standardize each effect by dividing by the appropriate between-subject standard deviation (SD), for example, the baseline SD of all subjects in a controlled trial. Although the magnitude of the resulting effects can be interpreted directly, differences in the SD between studies–reflecting different populations, different methods of sampling, or just sampling variation–will introduce heterogeneity that is unrelated to any real differences in the effect between studies. Josefsson et al. (2014) used standardization to cope with the various depression scales in different studies of the effect of exercise interventions on depression. In the six studies using the Beck depression inventory, the range in SD of baseline depression amounted to a factor of 7.5, which translates directly into artefactual heterogeneity. (The lowest SD is actually a standard

error in the original publication; Josefsson et al. apparently used it incorrectly to standardize.) The best approach for dealing with disparate psychometric scales, including visual-analogue and multi-level Likert scales, is to linearly re-scale all of them to a score with a minimum possible of 0 and a maximum possible of 100. The resulting score then represents an easily analyzed and interpreted percent of full range of the psychological construct.

Standardization probably contributed to heterogeneity in the effects of iron supplementation on serum iron concentration, blood hemoglobin concentration and VO2max in the meta-analysis of Burden et al. (2015). The authors provided no SD in the various studies to allow assessment of this issue or to convert the meta-analyzed standardized effect back into more meaningful units. By checking data in some of the original publications, I determined that much of the heterogeneity must have arisen from incorrect use of standard errors rather than SD for standardizing some effects. Standardization would also have contributed to heterogeneity in the meta-analysis that Salvador et al. (2016) conducted on the effect of ischemic preconditioning on exercise performance, if they had used the baseline standard deviation to standardize. Unfortunately, through misuse of the analysis program they effectively used the standard error of the change scores, which made the resulting standardized magnitudes meaningless.

Log transformation of effects expressed as factors followed by back-transformation to factors or percents is usually the best way to deal with physical performance and many other physiological measures. The decision between analysis of raw vs log-transformed effects hinges on which approach produces less heterogeneity. Peterson et al. (2011) used raw units (kg) for their meta-analysis of the effect of resistance exercise on lean body mass in aging adults, but if they had used log transformation, the apparently smaller effect in older adults would likely have disappeared or even reversed following back-transformation to percent units. Strangely, they provided irrelevant detail of a method of standardization using the SD of change scores, an error promulgated by the software package Comprehensive Meta-Analysis that again would have made the magnitudes meaningless. Dependent variables in the meta-analyses of Burden et al. (2015) and Salvador et al. (2016) were likely candidates for

log transformation, as was strength in the meta-analysis of Tomlinson et al. (2015) on the effects of supplementation with vitamin D. Chu et al. (2016) may have made the right choice of raw units of concentration in their meta-analysis of the effect of exercise on plasma zinc, but they did not show enough data from each study for me to assess whether the effects were more uniform as factors.

Effects on time-dependent events such as injury incidence are reported as ratios of odds, risks, incidence rates (hazards), or other rates (counts per some measure of exposure). When the risks are low–that is, <10% of the sample experience the event during the period of observation–all these ratios effectively have the same value, so they can be meta-analyzed and interpreted as risk ratios. When risks are higher, the hazard and other rate ratios coincide and are appropriate for interpreting and meta-analyzing factors affecting risk, whereas the odds ratio and the usual relative risk ratio increasingly overestimate and underestimate effects, respectively. Use of odds ratios by Wright et al. (2015) to meta-analyze risk factors in prospective studies of stress fractures in runners was therefore misguided, and they did not provide injury counts and sample sizes in each study to allow assessment of the extent to which their use of odds ratios introduced upward bias and artefactual heterogeneity in the risk factors they analyzed. (Anyone planning an injury meta-analysis should note that odds ratios in case-control studies are effectively hazard ratios, and that prevalence proportions should be meta-analyzed as odds ratios and converted back to proportion ratios for interpretation of clinical importance.)

Validity studies in which a practical measure is compared with a criterion provide several candidate measures for meta-analysis: the correlation coefficient, mean bias, bias at a predicted value, random error, and limits of agreement. The correlation coefficient is sensitive to the between-subject SD, so I would usually avoid it. If biases and errors are more uniform across studies when expressed as percents, they should be converted to factors and log-transformed before analysis. As an SD, the random error suffers from small-sample downward bias, a problem that is easily solved by expressing it as a variance (after any log-transformation), then taking the square root of the meta-analyzed mean. The variance also has a well-defined standard error,

which is needed for weighting the effects (see below). The bias in the SD is practically negligible for the usual sample sizes in validity studies, so no real harm was done when Coquart et al. (2016a) meta-analyzed SDs of the difference between criterion VO2max in a maximal test and VO2max predicted by submaximal tests. They then converted the meta-analyzed mean bias and mean random error into mean limits of agreement, and in a serious omission they provided no uncertainties (confidence limits) for any of these measures. Furthermore, they showed meta-analyzed random-error components as about ±4 ml.min^{-1}.kg^{-1}, an impossible outcome when the values in the individual studies ranged from ±10 to ±15 ml.min^{-1}.kg^{-1}. Understanding limits of agreement is evidently difficult enough without also having to consider their uncertainty, so this statistic should not be presented as the outcome of a meta-analysis of validity studies, or indeed of reliability studies.

Dealing with Standard Errors

A meta-analyzed effect is a weighted mean of effects, where the weighting factor is the inverse of the square of each effect's expected sampling variation, its standard error. Using a study-quality score as the weighting factor, as Hume et al. (2015) did in a meta-analysis of snow-sport injuries, is incorrect.

Depending on the design of the studies and the analysis package, the meta-analyst may input data or inferential statistics (p values or confidence limits) from each study without having to derive or impute the standard error for each effect. Exactly what was done needs to be stated, to satisfy readers that this step was performed correctly and to guide future meta-analyses. Coquart et al. (2016a) provided no inferential statistics or information about the standard errors for the two validity statistics they meta-analyzed, bias and random error. Tomlinson et al. (2015) input post-intervention means and SD into the meta-analysis software, when they should have input mean pre-post change scores and associated inferential statistics. The way Burden et al. (2015) combined pre and post scores is unclear, and Peterson et al. (2011) did not provide enough data from each study for me to check their analyses. Most of the other meta-analysts did (Chu et al., 2016; Josefsson et al., 2014; Salvador et al., 2016; Soomro et al., 2015; Wright et al., 2015), but only Chu et al. (2016) and Soomro et al. (2015) also provided adequate

documentation.

Accounting for Heterogeneity

Heterogeneity in a meta-analysis refers to real differences between effect magnitudes, which arise not from sampling variation but from moderation of the effect by differences between studies in subject characteristics, environmental factors, study design, measurement techniques, and/or method of analysis. The typical practice of testing for heterogeneity with the I^2 statistic is futile, because non-significance does not usually exclude the possibility of substantial heterogeneity, and neither the I^2 nor the related Q statistic properly represent the magnitude of heterogeneity (Higgins, 2008). The best statistic to deal with heterogeneity is the SD derived from a study-estimate random effect (often represented by τ), which should be included in *all* meta-analyses. This SD should be spelt out to readers as the typical difference in the true effects in different study settings. As such, it may be as important as the mean effect, in the same way that individual differences in the effect of an intervention may be as important as the mean. As with the SD representing individual differences, it should be doubled before it is interpreted against the magnitude thresholds normally used to interpret differences in means, or equivalently, the thresholds should be halved (Hopkins, 2015). The SD has its own uncertainty, which needs to be estimated, presented, and taken into account in the interpretation of its magnitude, and the analysis has to allow for negative estimates of its value and of its confidence limits.

Having shown that there *could* be substantial heterogeneity (i.e., the upper confidence limit is substantial), the meta-analyst should then try to explain it by performing separate meta-analyses of subgroups of studies or by performing a meta-regression with study characteristics, including mean subject characteristics, as predictors. The latter approach is preferable, especially when there are enough studies to allow for proper estimation and adjustment for mutual confounding with multiple predictors. These two approaches have been referred to disparagingly as *deluded* and *daft*, respectively, by authors who advocate what they call a *deft* approach: separate meta-analysis of the modifying effect of a given subject characteristic from each study (Fisher et al., 2017). Unfortunately, the deft approach is generally unrealizable: either the studies are done on relatively homogeneous groups of subjects (to obtain better precision for the kind of subject studied), or the authors do not report adequate information for the effect of the subject characteristic (e.g., only a p-value inequality), or they simply did not investigate the modifying effect of the subject characteristic. Meta-analysis with individual-participant data is probably the best way to account for modifying effects, but until such data become generally available, I recommend the daft approach, taking care to reduce bias by including potentially confounding study and mean subject characteristics in a single meta-regression. The daft approach still has the potential problem of so-called ecological bias, "whereby [modifying effects] at the aggregate level might not reflect the true [modifying effects] at the individual participant level" (Fisher et al., 2017). True, but analysis of individual participant data can itself produce biased estimates, such as attenuation by error of measurement, which is likely negligible with estimates based on study means. *Daft* is inappropriately dismissive.

Of the meta-analyses I reviewed for this article, only that of Soomro et al. (2015), on injury-prevention programs in team sports, included adequate assessment of heterogeneity and subgroup analyses. They did not present the random-effect SD in comprehensible units, nor did they present its uncertainty, but they did provide a *prediction interval* (Higgins, 2008; Higgins et al., 2009) representing the range of the true effect in 95% of study settings, akin to the reference range of a clinical test measure. Higgins et al. (2009) suggested that the prediction interval can be calculated by assuming the true values of the effect in different study settings have a t distribution centered on the meta-analyzed mean, with a variance given by the sum of the random-effect variance (τ^2) and error variance of the mean, and with degrees of freedom equal to the number of studies minus 2. These assumptions appear to me to be untenable, when there is sufficient uncertainty in τ^2 for its confidence interval to include negative values, which is the usual scenario with the kind of small-scale meta-analyses we see in our disciplines. Bootstrapping could be used to derive the lower and upper confidence limits for the lower and upper limits of the prediction interval (by setting negative values of τ^2 in the bootstrapped samples to zero), but easier statistics to understand would be the

proportions of study settings showing substantially positive, substantially negative, and trivial true effects. The meta-analytic models almost invariably involve the unrealistic assumption of a single estimate of τ^2 for different predicted means, so on balance I think it is best to go no further than interpreting τ and its confidence limits simply as the SD representing typical differences in the true effects between study settings.

Most other meta-analysts performed random-effect analyses and subgroup analyses, but any conclusions they based on the I^2 and Q statistics should be ignored. Josefsson et al. (2014) and Soomro et al. (2015) deserve a commendation for including study quality in a subgroup or moderator analysis. Only Peterson et al. (2011) performed meta-regression with multiple study characteristics. Burden et al. (2015) used only a fixed-effects model but included covariates when they thought the I^2 statistic indicated heterogeneity. Hume et al. (2015) performed fixed-effects meta-analyses without any consideration of heterogeneity, but they did perform subgroup analyses.

Coping with Repeated Measurement

A given study often provides several estimates of an effect that can be included in a meta-analysis, such as effects on males and females, or effects for different doses or time points. Such effects represent repeated measurement on the same study, so the usual meta-analytic mixed model with a single between-study random effect is not appropriate. Meta-analysts in the studies I reviewed attempted to cope with this problem either by treating the estimates as if they came from separate studies (Burden et al., 2015; Chu et al., 2016; Peterson et al., 2011; Salvador et al., 2016; Tomlinson et al., 2015), or by performing subgroup analyses that did not include repeated measurement (Coquart et al., 2016a). The problem with the former approach is that the resulting confidence interval for the overall mean effect is too narrow, while the resulting confidence intervals for any within-study moderators included in a meta-regression are wider than they need be. The problem with the latter approach is that the separately meta-analyzed effects in the subgroups cannot be compared inferentially, because they are not independent. The study of Coquart et al. (2016a) illustrates how wrong conclusions can be reached with an inappropriate analysis. They found similar meta-analyzed mean estimates of VO2max when they performed separate analyses for estimates predicted at perceived exertions of 19 and 20, but as you would expect, VO2max was substantially higher at the higher intensity in those studies where VO2max was predicted at both intensities, and I have little doubt that the right kind of meta-analysis would show that difference clearly.

The correct approach to including and comparing multiple within-study estimates is a repeated-measures meta-regression, achieved by including one or more covariates to account for and estimate the within-study effects, and by including one or more random effects additional to the usual between-study random effect to account for clustering of estimates within studies. I have only ever included a single additional random effect in meta-analyses (Carr et al., 2011; Vandenbogaerde and Hopkins, 2011; Weston et al., 2014), but in future I may use two random effects to account for within-study between-subject clustering (e.g., sex) and within-study within-subject clustering (e.g., multiple doses or time points).

Publication Bias and Outlier Studies

A pervasive tendency for only statistically significant effects to end up in print results in the overestimation of published effects, a phenomenon known as publication bias. Such bias was not an issue for the validity meta-analysis of Coquart et al. (2016a); five of the other meta-analysts did not mention the possibility of publication bias in their effects (Burden et al., 2015; Hume et al., 2015; Josefsson et al., 2014; Tomlinson et al., 2015; Wright et al., 2015) while four (Chu et al., 2016; Peterson et al., 2011; Salvador et al., 2016; Soomro et al., 2015) investigated asymmetry in the funnel-shaped plot of observed effects vs their standard errors, which is a sign of publication bias. There are two problems with this approach and corrections based on it: heterogeneity disrupts the funnel shape, thereby increasing the likelihood of false-negative and false-positive decisions about publication bias, and it does not take into account any heterogeneity explained in a meta-regression. A plot of the values of the study random-effect solution (effectively the study residuals) vs the study standard error solves these problems: publication bias manifests as a tendency for the residuals to be distributed non-uniformly

for studies with higher values of the standard error, and repeating the analysis after deleting all such studies reduces or removes the bias (see especially Carr et al., 2011; Vandenbogaerde and Hopkins, 2011). Standardization of the random-effect values converts them to z scores, which also allow for objective identification and elimination of outlier studies. Chu et al. (2016) and Peterson et al. (2011) investigated the change in the magnitude of the meta-analyzed effect following deletion of one or more studies, presumably checking for outliers; this kind of sensitivity analysis is pointless, because the change in magnitude will be smaller with a larger total number of meta-analyzed studies, and there is no associated rationale for eliminating studies. The researchers may have done this kind of analysis simply because it was available in the analysis package Comprehensive Meta-Analysis.

Interpreting Magnitudes

A shortcoming with several of the meta-analyses is inadequate attention to the clinical or practical importance of the meta-analyzed effects, let alone that of their moderators and heterogeneity. Coquart et al. (2016a) made no assessment of the implications of the magnitude of the meta-analyzed validity statistics for assessment of individual patients. Some authors apparently assumed that statistical significance automatically confers importance on the effect, without considering the magnitude of the observed effect or its confidence limits (Chu et al., 2016; Wright et al., 2015). Others used various scales to interpret standardized differences in means, without converting them back into real units to consider whether the standardized magnitude could represent an important clinical or practical effect in all or any populations (Burden et al., 2015; Salvador et al., 2016; Tomlinson et al., 2015). I support standardization for assessing differences or changes in means when there is no real-world scale, but the standardization should be done *after* the meta-analysis, using appropriately averaged between-subject SD from studies representing a population of interest. The average should be derived via variances weighted by degrees of freedom.

As already noted, the SD representing heterogeneity should be doubled (or squared for factor SD) before assessing its magnitude with the same scale as that for assessing the mean effect. Effects of moderators expressed as correlation coefficients (Salvador et al., 2016), "beta" coefficients (Burden et al., 2015; Peterson et al., 2011) and p values (Chu et al., 2016) do not communicate magnitude to the reader. Moderators representing numeric subject characteristics (e.g., mean age) that have been included as simple linear predictors should be evaluated for a difference in the characteristic equal to two between-subject SDs appropriately averaged from selected studies (again, via weighted variances). A suspected non-linear moderator can be coded as quantiles or other sub-group levels and evaluated accordingly.

Almost all the meta-analysts showed some reliance on p values to make conclusions, a practice that in my opinion is particularly inappropriate for meta-analyses. Inferences about the magnitude of all statistics should be based one way or another on the uncertainty represented by the magnitude of lower and upper confidence limits (Hopkins et al., 2009; Hopkins and Batterham, 2016).

Conclusion

I rate failure to use a random-effect meta-analysis and failure to properly account for heterogeneity in a random-effect meta-analysis as the most serious flaws, because heterogeneity combined with the mean effect determines a probabilistic range in the clinical or practical importance of the effect in a specific setting. As a researcher or practitioner, you should be cautious about implementing the findings of a meta-analysis lacking a full account of heterogeneity; use it primarily as a convenient reference list to find studies from settings similar to your own, and use these studies to draw your own conclusions about the magnitude of the effect in your setting.

You should also be skeptical about any meta-analyzed differences or changes in means based on standardization: there is a good chance the authors will have made major errors, and even when done correctly, standardization results in artefactual heterogeneity. Authors need to provide more documentation about these and the other error-prone aspects of meta-analysis I have identified here, if readers are to have more trust in the findings.

When I sent the first submitted version of this article to the authors of the meta-analyses for comment, one of them asked me to revise the article into a full meta-analysis of all recent meta-analyses. Such an article would represent a more

even-handed critique, given that these meta-analysts would likely find themselves in the company of the authors of most other recent meta-analyses. A longer article will be justified if the quality of meta-analyses in our subject areas does not improve in the next year or two.

References

Bonetti DL, Hopkins WG (2009). Sea-level exercise performance following adaptation to hypoxia. Sports Medicine 39, 107-127

Braakhuis AJ, Hopkins WG (2015). Impact of dietary antioxidants on sport performance: a review. Sports Medicine 45, 939-955

Burden RJ, Morton K, Richards T, Whyte GP, Pedlar CR (2015). Is iron treatment beneficial in, iron-deficient but non-anaemic (IDNA) endurance athletes? A meta-analysis. British Journal of Sports Medicine 49, 1389-1397

Carr AJ, Hopkins WG, Gore CJ (2011). Effects of acute alkalosis and acidosis on performance. Sports Medicine 41, 801-814

Cassar S, Misso ML, Hopkins WG, Shaw CS, Teede HJ, Stepto NK (2017). Insulin resistance in polycystic ovary syndrome: a systematic review and meta-analysis of euglycaemic-hyperinsulinaemic clamp studies. Human Reproduction 31, 2619–2631 (https://doi.org/2610.1093/humrep/dew2243)

Chu A, Petocz P, Samman S (2016). Immediate effects of aerobic exercise on plasma/serum zinc levels: a meta-analysis. Medicine and Science in Sports and Exercise 48, 726-733

Coquart JB, Tabben M, Farooq A, Tourny C, Eston R (2016a). Submaximal, perceptually regulated exercise testing predicts maximal oxygen uptake: a meta-analysis study. Sports Medicine 46, 885-897

Coquart JB, Tabben M, Tourny C, Chamari K, Eston RG (2016b). Author's Reply to Will G. Hopkins: "Submaximal, perceptually regulated exercise testing predicts maximal oxygen uptake: a meta-analysis study". Sports Medicine 46, 1197-1198

Fisher DJ, Carpenter JR, Morris TP, Freeman SC, Tierney JF (2017). Meta-analytical methods to identify who benefits most from treatments: daft, deluded, or deft approach? BMJ 356, j573

Higgins JPT (2008). Commentary: Heterogeneity in meta-analysis should be expected and appropriately quantified. International Journal of Epidemiology 37, 1158-1160

Higgins JPT, Thompson. SG, Spiegelhalter DJ (2009). A re-evaluation of random-effects meta-analysis. Jurnal of the Royal Statistical Society A 172, 137-159

Higgins JPT, Green S (2011). Cochrane Handbook for Systematic Reviews of Interventions Version 5.1.0 [updated March 2011]. Cochrane Collaboration. www.cochrane-handbook.org. Accessed 15 Jan 2016.

Hopkins WG (2004). An introduction to meta-analysis. Sportscience 8, 20-24

Hopkins WG, Marshall SW, Batterham AM, Hanin J (2009). Progressive statistics for studies in sports medicine and exercise science. Medicine and Science in Sports and Exercise 41, 3-12

Hopkins WG (2015). Individual responses made easy. Journal of Applied Physiology 118, 1444-1446

Hopkins WG (2016). Comment on: "Submaximal, perceptually regulated exercise testing predicts maximal oxygen uptake: a meta-analysis study". Sports Medicine 8, 1195-1196

Hopkins WG, Batterham AM (2016). Error rates, decisive outcomes and publication bias with several inferential methods. Sports Medicine 46, 1563-1573

Hume PA, Lorimer AV, Griffiths PC, Carlson I, Lamont M (2015). Recreational snow-sports injury risk factors and countermeasures: a meta-analysis review and Haddon matrix evaluation. Sports Medicine 45, 1175-1190

Josefsson T, Lindwall M, Archer T (2014). Physical exercise intervention in depressive disorders: Meta-analysis and systematic review. Scandinavian Journal of Medicine and Science in Sports 24, 259-272

McGrath LJ, Hopkins WG, Hinckson EA (2015). Associations of objectively measured built-environment attributes with youth moderate–vigorous physical activity: a systematic review and meta-analysis. Sports Medicine 45, 841-865

Peterson MD, Sen A, Gordon PM (2011). Influence of resistance exercise on lean body

mass in aging adults: a meta-analysis. Medicine and Science in Sports and Exercise 43, 249

Salvador A, De Aguiar R, Lisboa F, Pereira K, Cruz R, Caputo F (2016). Ischemic preconditioning and exercise performance: a systematic review and meta-analysis. International Journal of Sports Physiology and Performance 11, 4-14

Snowling NJ, Hopkins WG (2006). Effects of different modes of exercise training on glucose control and risk factors for complications in type 2 diabetic patients a meta-analysis. Diabetes Care 29, 2518-2527

Somerville VS, Braakhuis AJ, Hopkins WG (2016). Effect of flavonoids on upper respiratory tract infections and immune function: A systematic review and meta-analysis. Advances in Nutrition 7, 488-497

Soomro N, Sanders R, Hackett D, Hubka T, Ebrahimi S, Freeston J, Cobley S (2015). The eficacy of injury prevention programs in adolescent team sports: a meta-analysis. American Journal of Sports Medicine, https://doi.org/10.1177/0363546515618372

Tomlinson PB, Joseph C, Angioi M (2015). Effects of vitamin D supplementation on upper and lower body muscle strength levels in healthy individuals. A systematic review with meta-analysis. Journal of Science and Medicine in Sport 18, 575-580

Vandenbogaerde TJ, Hopkins WG (2011). Effects of acute carbohydrate supplementation on endurance performance. Sports Medicine 41, 773-792

Weston M, Taylor KL, Batterham AM, Hopkins WG (2014). Effects of low-volume high-intensity interval training (HIT) on fitness in adults: a meta-analysis of controlled and non-controlled trials. Sports Medicine 44, 1005-1017

Wright AA, Taylor JB, Ford KR, Siska L, Smoliga JM (2015). Risk factors associated with lower extremity stress fractures in runners: a systematic review with meta-analysis. British Journal of Sports Medicine 49, 1517-1523