HHS Public Access

Author manuscript

J Clin Epidemiol. Author manuscript; available in PMC 2015 December 21.

Published in final edited form as:

J Clin Epidemiol. 2015 February ; 68(2): 154–162. doi:10.1016/j.jclinepi.2014.09.003.

Statistical approaches to harmonize data on cognitive measures in systematic reviews are rarely reported

Lauren E. Griffith, Ph.D.[1], **Edwin van den Heuvel, Ph.D.**[2], **Isabel Fortier, Ph.D.**[3], **Nazmul Sohel, Ph.D.**[1], **Scott M. Hofer, Ph.D.**[4], **Hélène Payette, Ph.D.**[5], **Christina Wolfson, Ph.D**[3,6], **Sylvie Belleville, Ph.D.**[7], **Meghan Kenny, M.A.**[1], **Dany Doiron, M.P.P.**[3], and **Parminder Raina, Ph.D.**[1]

[1]Department of Clinical Epidemiology and Biostatistics, McMaster University, Hamilton, Ontario, Canada [2]Department of Epidemiology, University Medical Center Groningen, University of Groningen, Groningen, The Netherlands [3]Research Institute of the McGill University Health Centre, Montreal, Canada [4]Department of Psychology, University of Victoria, Victoria, British Columbia, Canada [5]Research Center on Aging, Health & Social Services Center - University Institute of Geriatrics of Sherbrooke and University of Sherbrooke, Sherbrooke, Canada [6]Department of Epidemiology and Biostatistics and Occupational Health, McGill University, Montreal, Canada [7]Research Center, Institut Universitaire de Gériatrie de Montréal and Psychology Department, Université de Montréal, Montreal, Canada

Abstract

Objectives—To identify statistical methods for harmonization which could be used in the context of summary data and individual participant data meta-analysis of cognitive measures.

Study Design and Setting—Environmental scan methods were used to conduct two reviews to identify: 1) studies that quantitatively combined data on cognition, and 2) general literature on statistical methods for data harmonization. Search results were rapidly screened to identify articles of relevance.

Results—All 33 meta-analyses combining cognition measures either restricted their analyses to a subset of studies using a common measure or combined standardized effect sizes across studies; none reported their harmonization steps prior to producing summary effects. In the second scan, three general classes of statistical harmonization models were identified: 1) standardization methods, 2) latent variable models, and 3) multiple imputation models; few publications compared methods.

Conclusions—Although it is an implicit part of conducting a meta-analysis or pooled analysis, the methods used to assess inferential equivalence of complex constructs are rarely reported or discussed. Progress in this area will be supported by guidelines for the conduct and reporting of

Corresponding author: Parminder Raina, McMaster University, Department of Clinical Epidemiology and Biostatistics, MIP-Suite 309A, 1280 Main Street West, Hamilton, Ontario, Canada, L8S 4K1, Telephone: 905-525-9140 ext. 22197, Fax: 905-522-7681, praina@mcmaster.ca.

The authors declare no financial conflicts of interest

the data harmonization and integration and by evaluating and developing statistical approaches to harmonization.

Keywords

harmonization; meta-analysis; cognition; individual participant data; data pooling

Introduction

Individual Participant Data (IPD) meta-analysis and pooled analysis, in which the original "raw" participant data from each study are brought together at a central location, have become increasingly popular methods to combine data from randomized controlled trials (RCT), as well as from observational studies [1–3]. IPD meta-analyses increase the power to detect differential treatment effects across individuals in an RCT and allow for adjustment of confounding factors in the meta-analysis of observational studies. The main advantage of IPD meta-analysis is that researchers can assess the influence of participant-level covariates on all collected outcomes and measured time points of interest, not all of which are reported in the literature [1]. IPD meta-analysis is particularly relevant to comparative effectiveness reviews (CERs) when conducting sub-group analyses and when combining evidence from RCTs and observational studies examining benefits, harms, adherence, or persistence [4]. Although IPD meta-analyses are relatively rare compared to aggregate data meta-analysis, there is an unprecedented amount of biological and phenotype data available to clinical, health, and social science researchers.[5] To maximize the utility of publicly-funded projects and increase the speed of scientific discovery there has been a worldwide push to leverage multiple data sources to explore important research questions.[6] However, combining IPD is scientifically and technically challenging as well as time consuming and costly to conduct [7].

Integration of IPD requires the generation of compatible (or harmonized) datasets across studies. Retrospective harmonization, the procedures aimed at achieving the comparability of previously collected data [8], is a fundamental step in conducting a scientifically rigorous meta-analysis. It is well recognized in conducting meta-analyses, one should limit integration to studies using clinically and methodologically compatible designs and methods. Researchers often use PICO (Patient problem, Intervention, Comparison, Outcome) to form specific research questions and facilitate targeted literature searches [9], but there are currently no standard guidelines proposed by systematic review organizations to determine whether or not data are similar enough to combine.

The similarity of data can be compromised on a number of levels, and even when the same measures are used, large operational differences can be found [10]. When considering combining data sets from across the world, these differences can be magnified. Regardless of whether the same measure has been used, differences are inevitably introduced due to language, culture, and method of administration. Furthermore, sampling strategies can be strikingly different in that results from different studies may reflect different segments of the population. Thus the process of harmonization is essential and those undertaking systematic

reviews must take these issues noted above into account prior to deriving common variables with the goal of creating an overall estimate of effect for a given intervention or exposure.

Harmonization of IPD is an iterative process composed of a series of steps that need to be undertaken and documented to ensure validity, reproducibility, and transparency of the harmonization process. The main steps to harmonization include: 1) defining the research question and selecting eligible studies, 2) evaluating the potential for harmonization, 3) processing study-specific data under a common format to generate the harmonized dataset, and 4) evaluating the success of the harmonization process [11].

Harmonizing variables into a common set of similar measures is often constructed with an "algorithmic" processing step [12]. For instance, the generation of compatible or inferentially equivalent information across studies can involve creating simple study-specific cut-points for a variable such as age, or by combining different response categories across studies to make them compatible. Algorithmic processing is straight forward and easy to implement when categories are sufficiently comparable, which may explain why the algorithms are typically never reported. There are occasions, however, when researchers would like to combine the results of constructs that are measured on different scales. For these scales it is less obvious how to equate one level on one scale to another level on another scale, which makes an algorithmic approach less appealing and particularly challenging. Cognitive ability is an example of such a complex construct.

Cognition is a complex process involving a large number of separate yet inter-related components. Cognition can be classified into different structures. The most psychometrically validated structure is the Cattell-Horn-Carroll (CHC) theory which identifies 10 broad stratum abilities comprising over 70 narrow abilities [13]. Evidence from a broad range of research provides support that different cognitive abilities have different construct validities (i.e., exhibit differential age-change functions; sensitivity to neurodegenerative disorders). There are no agreed upon pure measures of these cognitive components and many measures of cognition reported in primary studies may assess somewhat different underlying constructs or use tests that differ in their psychometric properties, creating a substantial challenge for those conducting systematic reviews.

The difficulty in combining cognitive measures is underscored in a systematic review of pharmacological treatment of dementia conducted by Raina, et al.[14] As the included studies used a wide variety of cognitive function measures as outcomes, the authors had to determine which measures could be statistically combined without any methodological guidance. In this report 20 different "general" scales were identified and only the most commonly reported MMSE[15] and the Alzheimer's Disease Assessment Scale,[16] were included in quantitative meta-analyses, therefore resulting in a loss of information.

To identify the landscape of methods currently being used to combine cognitive data in meta-analyses specifically, and to pool complex constructs across databases in general, we conducted two environmental scans. In an environmental scan the research question is not narrow, the search terms are quite broad, and single reviewers are involved in both consideration of eligibility of articles and in data extraction. In addition, the articles are not

reviewed for methodological quality in the usual sense of a systematic review, but methodological properties of the methods used are scrutinized. The scope is different from a full systematic review [17].

The purpose of our first scan was to identify what methods were used to quantitatively combine cognitive data in systematic reviews. Of particular interest was whether or not the cognitive measures combined in the meta-analysis differed across studies. Thus we could assess the current methods used by researchers to aggregate these different measures. The second, more general scan, identified statistical processing methods that have been used to create harmonized datasets for the purpose of meta-analysis or data pooling. This scan was not restricted to the harmonization of cognitive measures.

Methods

Studies that quantitatively combined cognitive measures

Search Strategy—The literature searches were conducted by a research librarian using Medline®, EMBASE®, Web of Science, and MathSciNet®. All databases were searched from January 1, 2001 to September 27, 2011. The search terms used were “cognition” and “meta-analysis”. The same search was undertaken using the Google search engine. The search results were screened to identify articles of relevance to this review. The references of relevant articles were also checked, and a search was conducted to identify more recent articles that cited the relevant articles.

Inclusion and Exclusion Criteria—Any study that quantitatively combined individual-level or aggregate-level data on cognitive measures and was published in English was eligible. Cognitive measures were defined as one or more standardized neuropsychological assessments (i.e., measuring global function, executive function, psychomotor speed, attention, memory, or intelligence).

Review Process—A single rater with training in epidemiology and psychology reviewed the titles and abstracts of all articles to identify which articles. The full-text was retrieved and reviewed for each article that passed the title and abstract screening. Study-level characteristics were extracted by one reviewer. These included: 1) populations, study design, and number of studies included in the meta-analysis, 2) intervention of interest (if appropriate), 3) inclusion criteria, 4) types of cognitive measures and domains measured, and 5) meta-analytic methods.

Studies using or describing statistical processing methods for harmonization

Search Strategy—A similar search strategy, using the same databases and years was used to identify literature describing statistical processing methods for harmonization. Identifying this literature, however, is challenging as there are no standard keywords or mesh terms used in bibliographic databases. The search terms were reviewed by a technical expert panel (TEP) comprised of experts in harmonization, meta-analysis, and neuropsychological research. The final search terms included: “individual patient data,” OR “IPD,” OR “pooling,” OR “multiple imputation,” OR “data harmonization,” OR “meta-analysis

methods." A similar search was performed using the Google search engine. The search results were screened, the references of relevant articles were checked, and a search for more recent articles that cited the articles already identified as being of interest was undertaken. These references were further supplemented by articles identified by the TEP to improve the comprehensiveness of the search.

Inclusion and Exclusion Criteria—Any study that reported statistical processing methods for the harmonization of study data was included. For the purpose of this review, harmonization was defined as "procedures aimed at achieving and improving the comparability of different surveys" [18]. This definition was adapted to include study designs other than surveys. For completeness, the search was supplemented with articles on the conduct and methodology of IPD meta-analysis, methods for evaluating equivalence (i.e., whether instruments measure the same construct or latent variable, latent trait, or factor across groups or over time), imputation methods, and examples of data harmonization.

Review Process—All identified articles underwent full text screening for relevance by at least two raters. Data extracted from the methodology articles included a description of the study population and design, statistical processing methods used, and the context in which it was used.

Results

Studies that quantitatively combined cognitive measures

There were 121 potential meta-analyses of cognition measures identified; of these, 47 abstracts passed the first level of screening and the full text articles were retrieved. The full text screening resulted in a total of 33 articles, which are summarized in Supplemental Table 1 [19–51]. All meta-analyses used aggregate data. Most (19 or 57.6%) of the meta-analyses included observational studies [33–51]; 14 (42.4%) were restricted to RCTs [20–32]. The populations included ranged from school-aged children to adults aged 55 and older. The primary focus of the studies varied greatly, but most used the cognitive tests as an outcome associated with a putative harmful agent (e.g., mobile phone electromagnetic fields) or positive factor (e.g., being an expert athlete), or after an intervention (e.g., comparing off-pump vs. on-pump coronary artery revascularization). The cognitive measures differed across the meta-analyses. Most meta-analyses included multiple instruments that measured different aspects of cognition, such as executive function, or psychomotor speed.

All of the authors of the aggregate data meta-analyses either restricted their analyses to a subset of studies utilizing a common cognitive measure or combined effect sizes across studies using different measures of cognition. In all cases, the cognitive measures were treated as continuous outcomes. The most common method of analysis was to combine standardized mean differences across studies. When the measures of cognition were consistent across studies or were comparable tests with a normalized scale, a weighted mean difference was used. Ten studies used meta-regression [20,22,25,31,35,36,41–43,51]; nine used a standardized effect size (e.g., Cohen's d, Hedges' g) as the dependent variable; [20,22,25,31,35,36,41,43,51] and one used a weighted mean difference of normalized comparable tests [42].

Studies using or describing statistical processing methods for harmonization

The scan of statistical methods used for harmonization resulted in 63 unique articles. Of the 63 articles, 53 (84.1%) met the inclusion criteria [2,3,8,18,52–100]. The 10 excluded articles are listed in online Appendix A. Seven of the 53 articles (13.2%) described the methods used for statistical harmonization [54,59,71,73,83,84,98] (Table 1). Ten articles (18.9 %) focused on the conduct of IPD meta-analysis [2,3,60,64,70,76,90,92,99,100] and 6 articles (11.3%) focused on IPD meta-analysis methodology [8,18,57,74,77,79]. Six articles (11.3%) reviewed imputation methods and the appropriateness of their use [58,66,86,91,94,95] and 2 articles (3.8%) described methods for evaluating equivalence of item functioning across study subgroups [62,96]. A summary of these 24 supplemental studies is in Supplemental Table 2. Finally, 22 articles (41.5%) reported the results of 16 unique statistical harmonization analyses undertaken in different contexts [52,53,55,56,61,63,65,67–69,72,75,78,80–82,85,87–89,93,97] (Supplemental Table 3).

Three general classes of statistical methods were identified in this scan. A summary of the assumptions and the application of this type of model are described in Table 2. One class used a simple linear- or z-transformation to create a common metric for combining constructs measured using different scales across datasets. An example of this class is in the Comparison of Longitudinal European Studies on Aging. When harmonization was deemed appropriate, some constructs were converted to a 0 to 1 scale by dividing a continuous score by its maximum score.

A second class of methods posits that a latent factor(s) underlies a set of measured items that can be modeled using linear factor analysis (if the items are continuous), two parameter logistic item response theory (if the items are binary), a polytomous Rasch model (if the items are ordinal), or moderated nonlinear factor analysis (MNFA) if there is a mix of binary, ordinal, and/or continuous items [51,68,95].. In each case, the first step is to construct a “conversion key” using one of the statistical models described above. This step models the relationship between the latent construct and the measured items. The second step uses the conversion key to convert the information onto a common scale. Measurement equivalence must then be assessed across samples [96].

The final class of methods, multiple imputation, is described by Burns, et al. [59]. The authors were interested in combining Mini-Mental State Examination (MMSE) scores with missing data across nine Australian longitudinal studies of aging. The MMSE score comprises 11 items, and the proportion of missing at least one MMSE item varied greatly by study and by wave of data collection. Furthermore, the data missingness was related to demographic characteristics, especially age and education. Burnsused a multiple imputation model with chained equations to impute missing MMSE item scores.

Supplemental Table 3 presents a summary of 22 publications from 16 data harmonization projects. Harmonization was often done by standardizing response options and determining whether questions were comparable across cohorts. For example, Minicuci, et al. [85] compared disability-free life expectancy using survey data collected in three populations. The authors used data on five questions assessing activities of daily living (ADL) that were common to all surveys. The response options for these questions were dichotomized to

create a common scale. Pluijm, et al. [87] similarly combined ADL data across six countries. There was overlap in the ADL items among the four items from the Katz ADL index; all four items were present in four of the six country surveys. In countries where the two items were not measured, the data for these were extrapolated from other “comparable” ADL items. Hot deck methods were used to impute values when one of the items was missing due to nonresponse.

Bath, et al. [53] harmonized cognitive data from the Longitudinal Aging Study Amsterdam (LASA) and the Nottingham Longitudinal Study on Activity and Ageing (NLSAA) by dividing each scale (Mini Mental State Exam and the Clifton Assessment Procedures for the Elderly by the maximum score for each instrument, MMSE/30 and CAPE/12, and combined them across studies.

Many of the studies used item response theory-based latent construct methods for analysis. van Buuren, et al. [97] used response conversion to create combinable international disability information, while Crane, et al. [61] used item response theory to co-calibrate cognitive scales. Both Curran, et al. [63] and Grimm, et al. [72] combined item response theory and growth curve models. Curran fit these models to data of developmental internalizing symptomatology, and Grimm examined the association between early behavioral and cognitive skills and later achievement. McArdle, et al. [82] used linear structural equation modeling with incomplete data to analyze repeated measures twin data to genetic and non-genetic factors associated with intellectual growth and change.

Schenker, et al. [89] combined clinical examination data with self-reported survey data from theNational Health and Nutrition Examination Survey. The National Health Interview Survey was larger and obtained a rich set of variables for use in multivariate analyses, but the study relied on self-report questions for the information on health conditions. Multiple imputation was used to properly reflect the sources of variability in subsequent analyses.

The Fibrinogen Studies Collaboration [69] combined data from 31 cohort studies using a two-stage approach. In the first stage partially and, where possible, fully adjusted estimates were obtained from each study, together with their standard errors. This method uses an imputation-type approach to address the issue of when studies included in an IPD meta-analysis include some, but not all, important confounding variables. . In the second stage, the study-specific estimates were combined.

Discussion

The environmental scan of aggregate data meta-analyses including cognitive measures revealed that all authors either restricted their analyses to a subset of common cognitive measures, or combined standardized effect sizes across studies. Although many of the meta-analyses reported study-specific information about the study populations, interventions, and cognitive outcomes, none reported formally exploring whether or not the cognitive measures should be combined or explicitly stated their harmonization steps prior to producing summary effects.

The environmental scan of statistical harmonization methods identified three general classes of methods. The first class uses a simple linear- or z-transformation to standardize the scale of measures to combine them across datasets. The second class of methods posits that there is a latent factor(s) that underlies a set of measured items that can be modeled, while the third class of methods was an "incomplete data" approach in which multiple imputation procedures or maximum likelihood estimation could be used to impute values for missing items. These items are then used to calculate a common scale that could be combined across studies, but imputation was typically not applied to items or scales that were missing by design, i.e., the items or scales were not intended to be part of the study.

Each method has strengths and weaknesses. The class of models that uses standardization methods has the most stringent assumptions (Table 2) which may not be appropriate when combining complex cognitive measures [101]. Dividing the scale by the maximum level transforms the scale to the same unit interval, but has essentially not changed the nature of the scales. The researchers must assume that the distribution of the standardized scale is mean and variance invariant. This means that it is assumed that the standardized scale is close to a normal distribution, in which only the mean and variance are important to investigate across studies. With scales for cognition though, ceiling effects may be present. When population characteristics change across studies, one study may demonstrate many more ceiling effects than other studies. When the scales have good item coverage at the boundaries (no or very little ceiling effects), standardization could be appropriate for harmonization. Of the latent construct approaches, the MNFA method proposed by Bauer is the most generalizable as it can accommodate different types of item data—binary, ordinal, or continuous—within a single model [54]. All of these approaches require that items can be "chained" together among studies, such that each study must have at least some items that overlap with another study. These bridge variables help standardize the latent variable across studies. The methods do assume that all the items give information about the same latent construct. This requires the same form of invariance that is implicitly used in standardizing scale, but this invariance is applied to the latent construct, which is more realistic than on the scale itself. Another potential limitation is that the methods require independent data within studies and the problems become much more complex for repeated measures in a longitudinal study. The authors using these methods tended to randomly choose one observation per person. Finally, latent variable models are much more complex to implement, in particular the general methods of MNFA, and may require sophisticated software or programming to be able to harmonize the scales. The latent construct approaches, are potentially the most promising and most general, because they try to capture the true information behind the observed measures, which is typically the goal of harmonization.

The final class of methods, based on multiple imputation[59], was used least frequently in the literature. The initial goal of multiple imputation is to provide valid estimates from incomplete data, which reflect the structure in the data, as well as the uncertainty about this structure. This type of model allows one to incorporate the factors that are related to missingness (e.g., demographic factors) in the imputation scheme. Additionally, missing items can also be imputed to complement studies. Then each study would contain the same set of variables and studies can be harmonized either through the use of algorithms or

through statistical processing like latent class models. This method requires that at least partly the same measures were included across studies, and that relationships between the variables that are used for imputation are consistent across studies. General issues around methods of imputation are reviewed by Peyre, et al. [86] and Spratt, et al. [94]. One philosophical discussion is whether variables that were missing by design can be imputed as well. Multiple imputation methods make use of the probability of being missing to generate or predict the missing values, but this probability is typically equal to one for variables that were missing by design. If this approach is considered appropriate it would also open a discussion on generating results in clinical trials for treatments that were not administered in the trial at all. For meta-analysis of mixed treatment comparisons “bridge treatments” could then play the role of bridge items.

In general, there was little focus in the literature on methods used to determine the inferential equivalence of variables prior to data integration through statistical processing. These harmonization steps may have, in fact, been conducted, but not reported. Granda, et al. [18] describe general approaches to harmonization and issues around determining cultural equivalence as a component of inferential equivalence. For example, Pluijm et al. [87] describe harmonizing measures of activities of daily living in older people across six countries. For some specific activities, questions used to collect data were similar, but there were cultural differences in meaning attached to the performance of the activities. For example, in Southern European countries older people receive help for cutting their toenails even if they do not have any difficulty in completing the task. The implication is that even when variables are standardized by such efforts as the Core Outcome Measures in Effectiveness Trials (COMET) Initiative [102], careful evaluation of the harmonization potential is required before processing data [103].

The environmental scans underscore the need for guidance on how to achieve harmonization and for the formal documentation of the harmonization process and the resulting methods used for statistical processing of complex constructs. Although it is an implicit part of conducting a meta-analysis or combined analysis, the methods used to assess inferential equivalence of complex constructs are rarely reported. In fact, the systematic review was complicated by the lack of standard search terms included in bibliographic databases around the harmonization process. The process of harmonization is essential and systematic reviewers must take these issues into account prior to deriving common variables that can be combined to create valid estimates of effect of a given intervention or exposure. Progress in this area will be supported by guidelines for the conduct and reporting of the data harmonization and integration to ensure the transparency and rigor of methods that will ultimately produce valid and reproducible harmonization results. Proposed recommendations for the conduct of harmonization for researchers undertaking IPD meta-analyses or data pooling and systematic review organizations are presented in figure 1.

Transparency in reporting harmonization methods, however, is just a first step. Methodological work is required to guide the choice of the most appropriate statistical processing approaches to integrate data from complex constructs in different contexts. It is clear that each method of statistical processing has specific assumptions, strengths and weaknesses. The appropriateness of the method used will be guided by the form of the

complex construct being harmonized. With the increase in IPD meta-analyses and push for pooled analyses across cohorts, the issue of harmonization and statistical processing of complex constructs will become increasingly important. Choosing the wrong approach or incorrectly specifying the model used to create derived variables might lead to bias or underestimate or overestimate within study variability, thus further methodologic work is required to understand the consequences of these choices to help guide researchers.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This manuscript is based on the methods research report Harmonization of Cognitive Measures in Individual Participant Data and Aggregate Data Meta-Analysis, funded by the Agency for Healthcare Research and Quality, United States Department of Health and Human Services under Contract No. 290 2007 10060 I. The authors are solely responsible for the content of the review. The opinions expressed herein do not necessarily reflect the opinions of the Agency for Healthcare Research and Quality. Lauren Griffith is supported by a CIHR New Investigators Award. Parminder Raina holds a Tier 1 Canada Research Chair in Geroscience and the Raymond and Margaret Labarge Chair in Research and Knowledge Application for Optimal Aging. Scott Hofer was supported by the National Institute on Aging, National Institutes of Health under Award Number P01AG043362.

Reference List

[1]. Oxman AD, Clarke MJ, Stewart LA. From science to practice - Metaanalyses using individual patient data are needed. JAMA. Sep 13; 1995 274(10):845–6. [PubMed: 7650811]

[2]. Riley RD, Lambert PC, Abo-Zaid G. Meta-analysis of individual participant data: Rationale, conduct, and reporting. BMJ. 2010; 340:c221. [PubMed: 20139215]

[3]. Blettner M, Sauerbrei W, Schlehofer B, Scheuchenpflug T, Friedenreich C. Traditional reviews, meta-analyses and pooled analyses in epidemiology. Int J Epidemiol. Feb; 1999 28(1):1–9. [PubMed: 10195657]

[4]. Slutsky J, Atkins D, Chang S, Sharp BAC. AHRQ Series Paper 1: Comparing medical interventions: AHRQ and the Effective Health-Care Program. J Clin Epidemiol. May; 2010 63(5):481–3. [PubMed: 18834715]

[5]. Khoury MJ. The case for a global human genome epidemiology initiative. Nat Genet. Oct; 2004 36(10):1027–8. [PubMed: 15454932]

[6]. Thompson A. Thinking big: large-scale collaborative research in observational epidemiology. Eur J Epidemiol. 2009; 24(12):727–31. doi: 10.1007/s10654-009-9412-1 [doi]. [PubMed: 19967428]

[7]. Griffith, L.; Shannon, H.; Wells, R.; Cole, D.; Hogg-Johnson, S.; Walter, S. The use of individual participant data (IPD) for examining heterogeneity in a meta-analysis of biomechanical workplace risk factors and low back pain. Fifth International Scientific Conference on Prevention of Work-Related Musculoskeletal Disorders; 2004. p. 337-338.

[8]. Granda, P.; Blasczyk, E. Guidelines for Best Practice in Cross-sectional Surveys. 2nd ed.. 2010. Data harmonization.

[9]. Schardt C, Adams MB, Owens T, Keitz S, Fontelo P. Utilization of the PICO framework to improve searching PubMed for clinical questions. BMC Med Inform Decis Mak. 2007; 7:16. doi: 1472-6947-7-16 [pii];10.1186/1472-6947-7-16 [doi]. Pubmed PMID: PMC1904193. [PubMed: 17573961]

[10]. Wiener JM, Hanley RJ, Clark R, Van Nostrand JF. Measuring the activities of daily living: Comparisons across national surveys. J Gerontol. 1990; 45(6):S229–37. [PubMed: 2146312]

[11]. Doiron D, Raina P, Ferretti V, L'Heureux F, Fortier I. Facilitating collaborative research: Implementing a platform supporting data harmonization and pooling. Norsk Epidemiologi. 2012; 21(2):221–4.

[12]. Fortier I, Burton PR, Robson PJ, Ferretti V, Little J, L'Heureux F, et al. Quality, quantity and harmony: The DataSHaPER approach to integrating data across bioclinical studies. Int J Epidemiol. Oct; 2010 39(5):1383–93. doi: dyq139 [pii];10.1093/ije/dyq139 [doi]. Pubmed PMID: PMC2972444. [PubMed: 20813861]

[13]. Flanagan, DP.; Ortiz, SO.; Alfonso, VC. Essentials of cross-battery assessment. 2nd ed.. Wiley; New York: 2007.

[14]. Raina P, Santaguida P, Ismaila A, Patterson C, Cowan D, Levine M, et al. Effectiveness of cholinesterase inhibitors and memantine for treating dementia: evidence review for a clinical practice guideline. Ann Intern Med. Mar 4; 2008 148(5):379–97. doi: 148/5/379 [pii]. [PubMed: 18316756]

[15]. Folstein MF, Folstein SE, McHugh PR. “Mini-mental state”. A practical method for grading the cognitive state of patients for the clinician. J Psychiatr Res. Nov; 1975 12(3):189–98. doi: 0022-3956(75)90026-6 [pii]. [PubMed: 1202204]

[16]. Mohs RC, Rosen WG, Davis KL. The Alzheimer's disease assessment scale: An instrument for assessing treatment efficacy. Psychopharmacol Bull. 1983; 19(3):448–50. [PubMed: 6635122]

[17]. Arksey H, O'Malley L. Scoping studies: towards a methodologic framework. International Journal of Social Research Methodology. 2005; 8(1):19–32.

[18]. Granda, P.; Wolf, C.; Hadorn, R. Harmonizing survey data. In: Harkness, JA.; Braun, M.; Edwards, B.; Johnson, TP.; Lyberg, L.; Mohler, PPH., et al., editors. Survey methods in multinational, multicultural and multiregional contexts. John Wiley & Sons; Hoboken, NJ: 2010. p. 315-32.

[19]. Angevaren M, Aufdemkampe G, Verhaar HJ, Aleman A, Vanhees L. Physical activity and enhanced fitness to improve cognitive function in older people without known cognitive impairment. Cochrane Database Syst Rev. 2008; (2):CD005381.

[20]. Eilander A, Gera T, Sachdev HS, Transler C, Van Der Knaap HCM, Kok FJ, et al. Multiple micronutrient supplementation for improving cognitive performance in children: Systematic review of randomized controlled trials. Am J Clin Nutr. 2010; 91(1):115–30. [PubMed: 19889823]

[21]. Falkingham M, Abdelhamid A, Curtis P, Fairweather-Tait S, Dye L, Hooper L. The effects of oral iron supplementation on cognition in older children and adults: A systematic review and meta-analysis. Nutrition Journal 9[4]. 2010 6-1-2012.

[22]. Guilera G, Pino O, Gomez-Benito J, Rojo JE. Antipsychotic effects on cognition in schizophrenia: A meta-analysis of randomised controlled trials. Eur J Psychiat. 2009; 23(2):77–89.

[23]. Hogervorst E, Bandelow S. Sex steroids to maintain cognitive function in women after the menopause: A meta-analyses of treatment trials. Maturitas. 2010; 66(1):56–71. [PubMed: 20202765]

[24]. Hogervorst E, Yaffe K, Richards M, Huppert FA. Hormone replacement therapy to maintain cognitive function in women with dementia. Cochrane Database Syst Rev. 2009; (1):CD003799. [PubMed: 19160224]

[25]. Li H, Li N, Li B, Li J, Wang P, Zhou T. Cognitive intervention for persons with mild cognitive impairment: A meta-analysis. Ageing Res Rev. 2011; 10:285–96. [PubMed: 21130185]

[26]. Karsdorp PA, Everaerd W, Kindt M, Mulder BJM. Psychological and cognitive functioning in children and adolescents with congenital heart disease: A meta-analysis. J Pediatr Psychol. 2007; 32(5):527–41. [PubMed: 17182669]

[27]. Lethaby A, Hogervorst E, Richards M, Yesufu A, Yaffe K. Hormone replacement therapy for cognitive function in postmenopausal women. Cochrane Database Syst Rev. 2008; (1):CD003122. [PubMed: 18254016]

[28]. Marasco SF, Sharwood LN, Abramson MJ. No improvement in neurocognitive outcomes after off-pump versus on-pump coronary revascularisation: A meta-analysis. Eur J Cardiothorac Surg. 2008; 33(6):961–70. [PubMed: 18424064]

[29]. Martin M, Clare L, Altgassen AM, Cameron MH, Zehnder F. Cognition-based interventions for healthy older people and people with mild cognitive impairment. Cochrane Database Syst Rev. 2011; 1:CD006220. [PubMed: 21249675]

[30]. Metternich B, Kosch D, Kriston L, Harter M, Hull M. The effects of nonpharmacological interventions on subjective memory complaints: A systematic review and meta-analysis. Psychother Psychosom. 2010; 79(1):6–19. [PubMed: 19887887]

[31]. Repantis D, Schlattmann P, Laisney O, Heuser I. Modafinil and methylphenidate for neuroenhancement in healthy individuals: A systematic review. Pharmacol Res. 2010; 62(3): 187–206. [PubMed: 20416377]

[32]. Woodward ND, Purdon SE, Meltzer HY, Zald DH. A meta-analysis of cognitive change with haloperidol in clinical trials of atypical antipsychotics: Dose effects and comparison to practice effects. Schizophr Res. 2007; 89(1–3):211–24. [PubMed: 17059880]

[33]. Campbell LK, Scaduto M, Sharp W, Dufton L, Van SD, Whitlock JA, et al. A meta-analysis of the neurocognitive sequelae of treatment for childhood acute lymphocytic leukemia. Pediatr Blood Canc. 2007; 49(1):65–73.

[34]. Goodman M, LaVerda N, Clarke C, Foster ED, Iannuzzi J, Mandel J. Neurobehavioural testing in workers occupationally exposed to lead: Systematic review and meta-analysis of publications. J Occup Environ Med. 2002; 59(4):217–23.

[35]. Grant I, Gonzalez R, Carey CL, Natarajan L, Wolfson T. Non-acute (residual) neurocognitive effects of cannabis use: A meta-analytic study. J Int Neuropsychol Soc. 2003; 9(5):679–89. [PubMed: 12901774]

[36]. Valentini E, Ferrara M, Presaghi F, De GL, Curcio G. Systematic review and meta-analysis of psychomotor effects of mobile phone electromagnetic fields. J Occup Environ Med. 2010; 67(10):708–16.

[37]. Wheaton P, Mathias JL, Vink R. Impact of early pharmacological treatment on cognitive and behavioral outcome after traumatic brain injury in adults: A meta-analysis. J Clin Psychopharmacol. 2009; 29(5):468–77. [PubMed: 19745647]

[38]. Barth A, Winker R, Ponocny-Seliger E, Mayrhofer W, Ponocny I, Sauter C, et al. A meta-analysis for neurobehavioural effects due to electromagnetic field exposure emitted by GSM mobile phones. J Occup Environ Med. 2008; 65(5):342–6.

[39]. Brands AMA, Biessels GJ, De Haan EHF, Kappelle LJ, Kessels RPC. The effects of type 1 diabetes on cognitive performance: A meta-analysis. Diabetes Care. 2005; 28(3):726–35. [PubMed: 15735218]

[40]. Sibley BA, Etnier JL. The relationship between physical activity and cognition in children: A meta-analysis. Pediatr Exerc Sci. 2003; 15(3):243–56.

[41]. Balint S, Czobor P, Komlosi S, Meszaros A, Simon V, Bitter I. Attention deficit hyperactivity disorder (ADHD): Gender- and age-related differences in neurocognition. Psychol Med. 2009; 39(8):1337–45. [PubMed: 18713489]

[42]. Bhutta AT, Cleves MA, Casey PH, Cradock MM, Anand KJS. Cognitive and behavioral outcomes of school-aged children who were born preterm: A meta-analysis. JAMA. 2002; 288(6):728–37. [PubMed: 12169077]

[43]. Bora E, Yucel M, Pantelis C. Cognitive endophenotypes of bipolar disorder: A meta-analysis of neuropsychological deficits in euthymic patients and their first-degree relatives. J Affect Disord. 2009; 113(1–2):1–20. [PubMed: 18684514]

[44]. Jansen CE, Miaskowski C, Dodd M, Dowling G, Kramer J. A metaanalysis of studies of the effects of cancer chemotherapy on various domains of cognitive function. Cancer. 2005; 104(10): 2222–33. [PubMed: 16206292]

[45]. Krabbendam L, Arts B, Van OJ, Aleman A. Cognitive functioning in patients with schizophrenia and bipolar disorder: A quantitative review. Schizophr Res. 2005; 80(2–3):137–49. [PubMed: 16183257]

[46]. McDermott LM, Ebmeier KP. A meta-analysis of depression severity and cognitive function. J Affect Disord. 2009; 119(1–3):1–8. [PubMed: 19428120]

[47]. Naguib JM, Kulinskaya E, Lomax CL, Garralda ME. Neuro-cognitive performance in children with type 1 diabetesa meta-analysis. J Pediatr Psychol. 2009; 34(3):271–82. [PubMed: 18635605]

[48]. Nieto RG, Xavier CF. A meta-analysis of neuropsychological functioning in patients with early onset schizophrenia and pediatric bipolar disorder. J Clin Child Adolesc Psychol. 2011; 40(2): 266–80. [PubMed: 21391023]

[49]. Quinn TJ, Gallacher J, Deary IJ, Lowe GD, Fenton C, Stott DJ. Association between circulating hemostatic measures and dementia or cognitive impairment: systematic review and meta-analyzes. J Thromb Haemost. Aug; 2011 9(8):1475–82. doi: 10.1111/j.1538-7836.2011.04403.x [doi]. [PubMed: 21676170]

[50]. Voss MW, Kramer AF, Basak C, Prakash RS, Roberts B. Are expert athletes 'expert' in the cognitive laboratory? A meta-analytic review of cognition and sport expertise. Appl Cognit Psychol. 2010; 24(6):812–26.

[51]. Zhang JP, Burdick KE, Lencz T, Malhotra AK. Meta-analysis of genetic variation in DTNBP1 and general cognitive ability. Biol Psychiatry. 2010; 68(12):1126–33. [PubMed: 21130223]

[52]. Anstey KJ, Byles JE, Luszcz MA, Mitchell P, Steel D, Booth H, et al. Cohort profile: The Dynamic Analyses to Optimize Ageing (DYNOPTA) project. Int J Epidemiol. Feb; 2010 39(1): 44–51. doi: dyn276 [pii];10.1093/ije/dyn276 [doi]. Pubmed PMID: PMC2817088. [PubMed: 19151373]

[53]. Bath P. The harmonisation of longitudinal data: A case study using data from cohort studies in The Netherlands and the United Kingdom. Ageing Soc. 2010; 30:1419–37.

[54]. Bauer DJ, Hussong AM. Psychometric approaches for developing commensurate measures across independent studies: Traditional and new models. Psychol Methods. Jun; 2009 14(2):101–25. doi: 2009-08072-001 [pii];10.1037/a0015583 [doi]. Pubmed PMID: PMC2780030. [PubMed: 19485624]

[55]. Beer-Borst S, Morabia A, Hercberg S, Vitek O, Bernstein MS, Galan P, et al. Obesity and other health determinants across Europe: the EURALIM project. J Epidemiol Community Health. Jun; 2000 54(6):424–30. Pubmed PMID: PMC1731700. [PubMed: 10818117]

[56]. Beer-Borst S, Hercberg S, Morabia A, Bernstein MS, Galan P, Galasso R, et al. Dietary patterns in six european populations: Results from EURALIM, a collaborative European data harmonization and information campaign. Eur J Clin Nutr. Mar; 2000 54(3):253–62. [PubMed: 10713749]

[57]. Bennett DA. Review of analytical methods for prospective cohort studies using time to event data: Single studies and implications for meta-analysis. Stat Methods Med Res. 2003; 12(4):297–319. [PubMed: 12939098]

[58]. Burgess S, Seaman S, Lawlor DA, Casas JP, Thompson SG. Missing data methods in Mendelian randomization studies with multiple instruments. Am J Epidemiol. Nov 1; 2011 174(9):1069–76. [PubMed: 21965185]

[59]. Burns RA, Butterworth P, Kiely KM, Bielak AA, Luszcz MA, Mitchell P, et al. Multiple imputation was an efficient method for harmonizing the Mini-Mental State Examination with missing item-level data. J Clin Epidemiol. Jul; 2011 64(7):787–93. doi: S0895-4356(10)00363-X [pii];10.1016/j.jclinepi.2010.10.011 [doi]. [PubMed: 21292440]

[60]. Cooper H, Patall EA. The relative benefits of meta-analysis conducted with individual participant data versus aggregated data. Psychol Methods. Jun; 2009 14(2):165–76. doi: 2009-08072-004 [pii];10.1037/a0015565 [doi]. [PubMed: 19485627]

[61]. Crane PK, Narasimhalu K, Gibbons LE, Mungas DM, Haneuse S, Larson EB, et al. Item response theory facilitated cocalibrating cognitive tests and reduced bias in estimated rates of decline. J Clin Epidemiol. Oct; 2008 61(10):1018–27. doi: S0895-4356(07)00437-4 [pii]; 10.1016/j.jclinepi.2007.11.011 [doi]. Pubmed PMID: PMC2762121. [PubMed: 18455909]

[62]. Crane PK, Narasimhalu K, Gibbons LE, Pedraza O, Mehta KM, Tang Y, et al. Composite scores for executive function items: demographic heterogeneity and relationships with quantitative magnetic resonance imaging. J Int Neuropsychol Soc. Sep; 2008 14(5):746–59. doi: S1355617708081162 [pii];10.1017/S1355617708081162 [doi]. Pubmed PMID: PMC2683684. [PubMed: 18764970]

[63]. Curran PJ, Hussong AM, Cai L, Huang W, Chassin L, Sher KJ, et al. Pooling data from multiple longitudinal studies: the role of item response theory in integrative data analysis. Dev Psychol. Mar; 2008 44(2):365–80. doi: 2008-02379-007 [pii];10.1037/0012-1649.44.2.365 [doi]. Pubmed PMID: PMC2894156. [PubMed: 18331129]

[64]. Curran PJ, Hussong AM. Integrative data analysis: the simultaneous analysis of multiple data sets. Psychol Methods. Jun; 2009 14(2):81–100. doi: 2009-08072-007 [pii];10.1037/a0015914 [doi]. Pubmed PMID: PMC2777640. [PubMed: 19485623]

[65]. Darby S, Hill D, Deo H, Auvinen A, Miguel BDJ, Baysson H, et al. Residential radon and lung cancer - detailed results of a collaborative analysis of individual data on 7148 persons with lung cancer and 14 208 persons without lung cancer from 13 epidemiologic studies in Europe. Scand J Psychol. Feb; 2006 32(Suppl(1)):1–83.

[66]. Donegan S, Williamson P, Gamble C, Tudur-Smith C. Indirect comparisons: A review of reporting and methodological quality. Plos One. Nov 10.2010 5(11):e11054. [PubMed: 21085712]

[67]. Duncan GJ, Dowsett CJ, Claessens A, Magnuson K, Huston AC, Klebanov P, et al. School readiness and later achievement. Dev Psychol. Nov; 2007 43(6):1428–46. doi: 2007-16709-012 [pii];10.1037/0012-1649.43.6.1428 [doi]. [PubMed: 18020822]

[68]. Esteve A, Sobek M. Challenges and methods of international census harmonization. Hist Meth. 2003; 36:66–79.

[69]. Correcting for multivariate measurement error by regression calibration in meta-analyses of epidemiological studies. Stat Med. 2009; 28(7):1067–92. [PubMed: 19222086]

[70]. Friedenreich CM. Methods for pooled analyses of epidemiologic studies. Epidemiol. Jul; 1993 4(4):295–302.

[71]. Gorsuch R. New procedure for extension analysis in exploratory factor analysis. EPM. 1997; 57(5):725–40.

[72]. Grimm KJ, Steele JS, Mashburn AJ, Burchinal M, Pianta RC. Early behavioral associations of achievement trajectories. Dev Psychol. Sep; 2010 46(5):976–83. doi: 2010-17955-002 [pii]; 10.1037/a0018878 [doi]. [PubMed: 20822216]

[73]. Gross AL, Inouye SK, Rebok GW, Brandt J, Crane PK, Parisi JM, et al. Parallel but not equivalent: Challenges and solutions for repeated assessment of cognition over time. J Clin Exp Neuropsychol. Aug; 2012 34(7):758–72. doi: 10.1080/13803395.2012.681628 [doi]. [PubMed: 22540849]

[74]. Hofer SM, Piccinin AM. Integrative data analysis through coordination of measurement and analysis protocol across independent longitudinal studies. Psychol Methods. Jun; 2009 14(2): 150–64. doi: 2009-08072-003 [pii];10.1037/a0015566 [doi]. Pubmed PMID: PMC2773828. [PubMed: 19485626]

[75]. Hopman-Rock M, van BS, De Kleijn-De VM. Polytomous Rasch analysis as a tool for revision of the severity of disability code of the ICIDH. Disabil Rehabil. May 20; 2000 22(8):363–71. [PubMed: 10896097]

[76]. Ioannidis JPA, Rosenberg PS, Goedert JJ, O'Brien TR. Commentary: Meta-analysis of individual participants' data in genetic epidemiology. Am J Epidemiol. Aug 1; 2002 156(3):204–10. [PubMed: 12142254]

[77]. Jones AP, Riley RD, Williamson PR, Whitehead A. Meta-analysis of individual patient data versus aggregate data from longitudinal clinical trials. Clin Trials. Feb; 2009 6(1):16–27. doi: 6/1/16 [pii];10.1177/1740774508100984 [doi]. [PubMed: 19254930]

[78]. Khachaturian AS, Sabbagh M. Commentary on “Developing a national strategy to prevent dementia: Leon Thal Symposium 2009.” Creating a national database for successful aging. Alzheimers Dement. Mar; 2010 6(2):132–4. doi: S1552-5260(10)00017-8 [pii];10.1016/j.jalz. 2010.01.012 [doi]. [PubMed: 20298973]

[79]. Mathew T, Nordstrom K. Comparison of one-step and two-step meta-analysis models using individual patient data. Biom J. Apr; 2010 52(2):271–87. doi: 10.1002/bimj.200900143 [doi]. [PubMed: 20349448]

[80]. McArdle, JJ.; Nesselroade, JR. Using multivariate data to structure developmental change. In: Cohen, SH.; Reese, HW., editors. Life-span developmental psychology: Methodological contributions. Lawrence Erlbaum Associates, Inc; Hillsdale, NJ, England: 1994. p. 223-67.

[81]. McArdle, JJ.; Prescott, CA. Contemporary models for the biometric genetic analysis of intellectual abilities. In: Flanagan, DP.; Genshaft, JL.; Harrison, PL., editors. Contemporary

intellectual assessment: Theories, tests, and issues. Guilford Press; New York, NY, US: 1997. p. 403-36.

[82]. McArdle JJ, Prescott CA, Hamagami F, Horn JL. A contemporary method for developmental-genetic analyses of age changes in intellectual abilities. Dev Neuropsychol. 1998; 14(1):69–114.

[83]. McArdle JJ, Grimm KJ, Hamagami F, Bowles RP, Meredith W. Modeling life-span growth curves of cognition using longitudinal data with multiple samples and changing scales of measurement. Psychol Methods. Jun; 2009 14(2):126–49. doi: 2009-08072-002 [pii];10.1037/a0015857 [doi]. Pubmed PMID: PMC2831479. [PubMed: 19485625]

[84]. Minicuci N, Noale M, Bardage C, Blumstein T, Deeg DJ, Gindin J, et al. Cross-national determinants of quality of life from six longitudinal studies on aging: The CLESA project. Aging Clin Exp Res. Jun; 2003 15(3):187–202. [PubMed: 14582681]

[85]. Minicuci N, Noale M, Leon Diaz EM, Gomez LM, Andreotti A, Mutafova M. Disability-free life expectancy: A cross-national comparison among Bulgarian, Italian, and Latin American older population. J Aging Health. Jun; 2011 23(4):629–81. doi: 0898264310390940 [pii];10.1177/0898264310390940 [doi]. [PubMed: 21220352]

[86]. Peyre H, Leplege A, Coste J. Missing data methods for dealing with missing items in quality of life questionnaires. A comparison by simulation of personal mean score, full information maximum likelihood, multiple imputation, and hot deck techniques applied to the SF-36 in the French 2003 decennial health survey. Qual Life Res. Mar; 2011 20(2):287–300. doi: 10.1007/s11136-010-9740-3 [doi]. [PubMed: 20882358]

[87]. Pluijm SM, Bardage C, Nikula S, Blumstein T, Jylha M, Minicuci N, et al. A harmonized measure of activities of daily living was a reliable and valid instrument for comparing disability in older people across countries. J Clin Epidemiol. Oct; 2005 58(10):1015–23. doi: S0895-4356(05)00144-7 [pii];10.1016/j.jclinepi.2005.01.017 [doi]. [PubMed: 16168347]

[88]. Ruggles S, King ML, Levison D, McCaa R, Sobek M. IPUMS-International. Hist Meth. 2003; 36:60–5.

[89]. Schenker N, Raghunathan TE. Combining information from multiple surveys to enhance estimation of measures of health. Stat Med. Apr 15; 2007 26(8):1802–11. doi: 10.1002/sim.2801 [doi]. [PubMed: 17278184]

[90]. Schmid CH, Landa M, Jafar TH, Giatras I, Karim T, Reddy M, et al. Constructing a database of individual clinical trials for longitudinal analysis. Control Clin Trials. Jun; 2003 24(3):324–40. doi: S0197245602003197 [pii]. [PubMed: 12757997]

[91]. Siddique J, Crespi CM, Gibbons RD, Green BL. Using latent variable modeling and multiple imputation to calibrate rater bias in diagnosis assessment. Stat Med. Jan 30; 2011 30(2):160–74. [PubMed: 21204122]

[92]. Simmonds MC, Higgins JP, Stewart LA, Tierney JF, Clarke MJ, Thompson SG. Meta-analysis of individual patient data from randomized trials: A review of methods used in practice. Clin Trials. 2005; 2(3):209–17. [PubMed: 16279144]

[93]. Slimani N, Kaaks R, Ferrari P, Casagrande C, Clavel-Chapelon F, Lotze G, et al. European Prospective Investigation into Cancer and Nutrition (EPIC) calibration study: Rationale, design and population characteristics. Public Health Nutr. Dec; 2002 5(6B):1125–45. doi: 10.1079/PHN2002395 [doi];S1368980002001362 [pii]. [PubMed: 12639223]

[94]. Spratt M, Carpenter J, Sterne JA, Carlin JB, Heron J, Henderson J, et al. Strategies for multiple imputation in longitudinal studies. Am J Epidemiol. Aug 15; 2010 172(4):478–87. doi: kwq137 [pii];10.1093/aje/kwq137 [doi]. [PubMed: 20616200]

[95]. Sterne JA, White IR, Carlin JB, Spratt M, Royston P, Kenward MG, et al. Multiple imputation for missing data in epidemiological and clinical research: Potential and pitfalls. BMJ. 2009; 338:b2393. Pubmed PMID: PMC2714692. [PubMed: 19564179]

[96]. Teresi JA, Ocepek-Welikson K, Kleinman M, Cook KF, Crane PK, Gibbons LE, et al. Evaluating measurement equivalence using the item response theory log-likelihood ratio (IRTLR) method to assess differential item functioning (DIF): Applications (with illustrations) to measures of physical functioning ability and general distress. Qual Life Res. 2007; 16(Suppl 1):43–68. doi: 10.1007/s11136-007-9186-4 [doi]. [PubMed: 17484039]

[97]. van Buuren S, Eyres S, Tennant A, Hopman-Rock M. Assessing comparability of dressing disability in different countries by response conversion. Eur J Public Health. Sep; 2003 13(3 Suppl):15–9. [PubMed: 14533743]

[98]. van Buuren S, Eyres S, Tennant A, Hopman-Rock M. Improving comparability of existing data by response conversion. J Off Stat. 2005; 21(1):53–72.

[99]. van der Steen JT, Kruse RL, Szafara KL, Mehr DR, van der Wal G, Ribbe MW, et al. Benefits and pitfalls of pooling datasets from comparable observational studies: Combining US and Dutch nursing home studies. Palliat Med. Sep; 2008 22(6):750–9. doi: 22/6/750 [pii]; 10.1177/0269216308094102 [doi]. [PubMed: 18715975]

[100]. van Walraven C. Individual patient meta-analysis--rewards and challenges. J Clin Epidemiol. Mar; 2010 63(3):235–7. doi: S0895-4356(09)00110-3 [pii];10.1016/j.jclinepi.2009.04.001 [doi]. [PubMed: 19595573]

[101]. Meredith, W. Notes on factorial invariance; Psychometrika. 1964. p. 177-85.doi: http://dx.doi.org/10.1007/BF02289699

[102]. Williamson P, Altman D, Blazeby J, Clarke M, Gargon E. Driving up the quality and relevance of research through the use of agreed core outcomes. J Health Serv Res Pol. 2012; 17(1):1–2.

[103]. Fortier I, Doiron D, Burton P, Raina P. Invited commentary: Consolidating data harmonization--how to obtain quality and applicability? Am J Epidemiol. Aug 1; 2011 174(3):261–4. doi: kwr194 [pii];10.1093/aje/kwr194 [doi]. [PubMed: 21749975]

What is new?

- *Clinicians, patients and policymakers* would benefit from making optimal use of all available research data, contingent on quality, to better understand disease processes and provide their best estimate of the impact of interventions.
- Combining data from measurements of complex constructs, such as cognition, requires a rigorous approach as well as specialized methods of harmonization.
- Although several meta-analyses combining cognitive measures have been published, none explicitly described their methods of harmonization.
- Our literature scan identifies several statistical approaches to processing harmonized data used in the context of meta-analysis and data pooling, but few studies compared methods.
- Progress in this area will be supported by guidelines for the conduct and reporting of the data harmonization and integration process, and by evaluating and developing statistical approaches to harmonization.

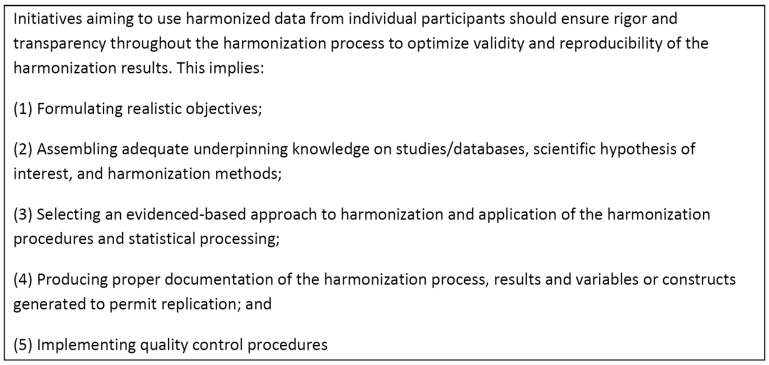

Figure 1.

Proposed harmonization recommendations for researchers and systematic review organization conducting IPD meta-analyses and data pooling

J Clin Epidemiol. Author manuscript; available in PMC 2015 December 21.

Table 1

Summary articles describing statistical methods for data harmonization

Study	Method	Context	Description	Pro	Con
Bauer, DJ 2009 [51]	Linear factor analysis (LFA)	Different psychometric methods for developing commensurate measures in the context of integrative data analysis (the simultaneous analysis of data obtained from two or more independent studies) were compared	**LFA** • A latent factor(s) is/are posited to underlie a set of observed, continuous variables • Must test the invariance of the latent factor(s) when comparing across groups	• Methodology well known • As long as there is a subset of invariant items, the factor can be combined across studies • Can include noncommon items in analyses • Even if no items are common to all studies, can still be conducted if studies can be “chained” together. For example, if item sets A and B are available in study 1, item sets B and C are available in study 2 and item sets C and D are available in study 3	• The validity of the results is dependent on the method used to identify the model. For example, the use of the reference item method (i.e., constraining the intercept and loading of one item to 0 and 1 in both groups) implicitly assumes the reference item is invariant across groups. Another option is to set the factor mean and variance to 0 and 1 in one group then estimate factor mean and variance in the other group while placing equality constraints on one or more item intercepts and loadings. This second procedure only works well when the number of noninvariant items is small relative to the number of invariant items • Requires continuous indicators • Requires more than one common item to measure equivalence • Sample units must be independent of one another (i.e., no repeated measures)
	Two-parameter logistic (2-PL) using Item response theory (IRT)		**2-PL IRT** • Assumes a single latent trait underlies a set of	• Widely used methodology (especially in testing)	• Requires binary indicators. For example, it could be used to measures

Study	Method	Context	Description	Pro	Con
			binary responses. Some items may be more difficult than others and some items may be more strongly related to the latent trait than It is typically assumed the latent trait has a standard normal distribution	• As long as there is a subset of invariant items, the factors can be combined across studies • Can include noncommon items in analyses Even if no items are common to all studies, can still be conducted if studies can be "chained" together. For example, if item sets A and B are available in study 1, item sets B and C are available in study 2 and item sets C and D are available in study 3	dimensions of psychopathology by using a set of symptom indicators • Requires more than one common item to measure equivalence • Sample units must be independent of one another (i.e., no repeated measures)
	Moderated nonlinear factor analysis (MNLFA)		**MNLFA** Uses generalized factor analysis models that can incorporate binary, ordinal, continuous and count items. The key parameters of the factor model (i.e., the indicator intercepts, factor loadings, and factor mean and variance) are permitted to vary as a function of one or more exogenous variables	• Can accommodate different scale types among items	• Requires more than one common item to measure equivalence Sample units must be independent of one another (i.e., no repeated measures)
Burns, RA. 2011 [56]	Multiple Imputation	Combining MMSE scores with missing data across 9 Australian longitudinal studies of aging (Dynamic Analyses to Optimize Aging [DYNOPTA] project) • Participants missing at least one MMSE item varied greatly by contributing study and wave of data collection [range 0.5%–95%] • Missing information related to demographic characteristics, especially age and education	MI used imputer model in multiple imputation with chained equations (MICE). The model included: gender, years of education, study and study interactions. Created 5 imputation datasets and took the average of the 5 imputed plausible values Software: MICE add-on to STATA version 10	• Does not require special software to run MICE subsequent analyses (i.e., one final dataset with average values) • Compared to single imputation methods, MI accounts for uncertainty in the missing value by using a set of plausible values	• Requires the same measures across studies • Requires item level data
Gorsuch, R. 1997 [68]	Extension analysis in exploratory factor analysis	In exploratory factor analysis extension analysis refers to computing the relationship among common factors to	A factor analysis is conducted on the core variables. The correlations between the core variables and the extension variables to estimate the	• Widely used methodology	• Requires the core set of variables to be common to all studies

Study	Method	Context	Description	Pro	Con
		variables that were not included in the factor analysis. For example, this would be used in a situation where a factor analysis would include proven items but now new experimental	[illegible]		• Requires the [illegible] • Only appropriate for continuous variables
McArdle, JJ. 2009 [80]	IRT combined with latent growth/decline curve modeling	This method was applied to longitudinal data from different cognitive test batteries to examine how to best model changes in cognitive constructs over a life span. The data come from 3 studies on intellectual abilities (Berkeley Growth Study (BGS), Guidance–Control Study (GCS), and Bradway–McArdle Longitudinal (BML) Study). The cognitive constructs measured were vocabulary and memory using 8 different intelligence test batteries. Although the tests were common items among the tests, different tests were between studies and over time within studies.	The authors consider several techniques for linkage across measurement scales and across multiple groups and fit a unidimensional Rasch model to item responses and a latent curve model together with changing latent scores over age and groups. The latent growth/decline curve model had a separate within-time measurement equation and over-time functional change equation. Because some items used in these analyses have graded outcome scores (i.e., 0, 1, or 2), a partial credit model was used for the IRT model. The parameters of both IRT and latent curve models were simultaneously estimated based on a joint model likelihood approach	• Can be used without complete overlap of items • Items can be linked by data (i.e., bridge studies and bridge items), by assuming equivalence, or by a combination of the two • Allows for a varying number of data points per person • Allows instruments to change over time within an individual • Method allows one to separate out differences in scales over time from changes in constructs over time	• Requires overlapping information across studies
Minicuci, N. 2003 [81]	Multiple methods including recategorization and z-score transformations	Constructed a harmonized measures using data from six countries [Finland, Italy, the Netherlands, Spain, Sweden and Israel] contributing data to the Comparison of Longitudinal European Studies on Aging (CLESA) Study. The first goal of the study was to create a common data base (CDB) with a framework to include behavioral, social, psychological, and health status measures. A common measure was created if at least 3 countries had measured the construct of interest.	Harmonization guidelines were developed for each type of variable. When harmonization was deemed appropriate, the most common methods for harmonization were to recategorize variables into a common set of response option and to create a common scale, e.g., 0–1, by dividing a continuous score by its maximum score Another related method of conversion is to create z-scores for each construct by subtracting the overall mean and dividing the raw score by the standard deviation.	• Relatively simple and does not require specialized statistical software • Does not require common items across studies	• Does not take into account the difference in distributions/ variability across populations • Assumes the underlying constructs are the same and measured equally well across populations
Gross, A.L. 2012 [70]	Mean, linear, and percentile transformations	Constructed mean, linear and percentile equating using data from two large-scale, multi-site cohorts: the Advanced Cognitive Training for Independent and Vital Elderly (ACTIVE) and the Alzheimer's Disease Neuroimaging Initiative (ADNI).	Used a two stage approach. In the first stage, an equating sample was selected from which to collect necessary characteristics of test distributions and derive the equating algorithm. In the second stage, equating algorithms were applied to the full study sample in	• Can be used with longitudinal data • Can be used to adjust	• Assumes that the population producing responses on different scaled tests at each time point have the same underlying ability

Study	Method	Context	Description	Pro	Con
		ACTIVE is a longitudinal randomized trial of cognitive training in cognitively intact, community dwelling adults age 65 and older ADNI was a five-year observational cohort study ACTIVE is a longitudinal randomized trial of cognitive training in cognitively intact, community dwelling adults age 65 and older ADNI was a five-year observational cohort study ACTIVE is a longitudinal randomized trial of cognitive training in cognitively intact, community dwelling adults age 65 and older ADNI was a five-year observational cohort study ACTIVE is a longitudinal randomized trial of cognitive training in cognitively intact, community dwelling adults age 65 and older ADNI was a five-year observational cohort study ACTIVE is a longitudinal randomized trial of cognitive training in cognitively intact, community dwelling adults age 65 and older ADNI was a five-year observational cohort study ACTIVE is a longitudinal randomized trial of cognitive training in cognitively intact, community dwelling adults age 65 and older ADNI was a five-year observational cohort study ACTIVE is a longitudinal randomized trial of cognitive training in cognitively intact, community dwelling adults age 65 and older ADNI was a five-year observational cohort study ACTIVE is a longitudinal randomized trial of cognitive training in cognitively intact, community dwelling adults age 65 and older ADNI was a five-year observational cohort study ACTIVE is a longitudinal randomized trial of cognitive training in cognitively intact, community dwelling adults age 65 and older ADNI was a five-year observational cohort study ACTIVE is a longitudinal randomized trial of cognitive training in cognitively intact, community dwelling adults age 65 and older ADNI was a five-year observational cohort study ACTIVE is a longitudinal randomized trial of cognitive training in cognitively intact, community dwelling adults age 65 and older ADNI was a five-year observational cohort study ACTIVE is a longitudinal randomized trial of cognitive training in cognitively intact, community dwelling adults age 65 and older ADNI was a five-year observational cohort study ACTIVE is a longitudinal randomized trial of cognitive training in cognitively intact, community dwelling adults age 65 and older ADNI was a five-year observational cohort study ACTIVE is a longitudinal randomized trial of cognitive training in cognitively intact, community dwelling adults age 65 and older ADNI was a five-year observational cohort study ACTIVE is a longitudinal randomized trial of cognitive training in cognitively intact, community dwelling adults age 65 and older ADNI was a five-year observational cohort study ACTIVE is a longitudinal randomized trial of cognitive training in cognitively intact, community dwelling adults age 65 and older ADNI was a five-year observational cohort study ACTIVE is a longitudinal randomized trial of cognitive training in cognitively intact, community dwelling adults age 65 and older ADNI was a five-year observational cohort study			
van Buuren, S. 2005 [95]	Response conversion (RC)	This method was applied to binary and original data measuring walking disability measured across 10 European countries.	RC is a two-step method. The first step is to construct a conversion key using a statistical model (e.g., polytomous Rasch model). This step models the relationship between the common scale and the measured items. The second step uses a conversion key to convert information onto a common scale	• Can be used without complete overlap of items • Items can be linked by data (i.e., bridge studies and bridge items), by assuming equivalence, or by a combination of the two	• Requires overlapping information across studies

Abbreviations: 2PL-IRT = two-parameter logistic using item response theory; BGS = Berkeley Growth Study; BML = Bradley-McArdle Intelligence Scale; CDB = common database; CLESA = Comparison of Longitudinal European Studies on Aging; DYNOPTA = Dynamic Analyses to Optimizing Ageing; GCS = Guidance-Control Study; IRT = item response theory; LFA = linear factor analysis; MI = multiple imputation; MICE = multiple imputation with chained equations; MMSE = Mini Mental State Examination; MNLFA = moderated nonlinear factor analysis; RC = response conversion; SB = Stanford-Binet; WAIS (-R) = Wechsler Adult Intelligence Scale (Revised); WB = Wechsler-Bellevue; WJ = Woodcock Johnson Psycho Educational Battery-Revised

Table 2

Assumptions for the different classes of statistical harmonization methods

Method	Assumptions	How can it be applied
Standardization Methods 6 studies used this class of methods, e.g., Minicuci, N. 2003 [81]	▪ Scale scores have an underlying normal distribution ▪ The scales have a similar distribution (i.e., being in the 5th percentile of one scale is equivalent to being in the 5th percentile of another)	▪ Can be applied in most situations with continuous variables and does not require specialized software ▪ Does not require common items across studies ▪ Need to transform back to a chosen scale(s) for interpretation
Item Response Theory Latent Variable Model 15 studies used this class of methods, e.g., Van Buuren, S. 2005; [95] Bauer, DJ. 2009; [51] McArdle, J. 2009 [80]	▪ Underlying constructs are unidimensional ▪ Some items must be common across datasets or at least can be "chained" together ▪ The items are equally discriminating (only for IP and Rasch models) ▪ Factorial invariance If repeated measures: ▪ Item difficulty is invariant with respect to time or age ▪ Item discrimination does not change across time or age	▪ Can be applied to continuous, binary and ordinal data but requires some specialized software ▪ Can accommodate different scale types among items ▪ However can be extended to include longitudinal data as per McArdle, et al. by integrating IRT and latent curve modeling using a joint model likelihood approach
Missing data by design with multiple imputation 3 studies used this class of methods, e.g., Burns, RA. 2011 [56]	▪ Missingness is assumed to be at random (i.e., MAR) ▪ Some items must be common across datasets or at least can be "chained" together	▪ Can be applied to continuous, binary and ordinal data but requires some specialized software and multiple datasets ▪ Can accommodate different scale types among items ▪ Can be used if scales are not unidimensional

Abbreviations: IRT = item response theory; MAR = missing at random