Toxicology Letters 206 (2011) 144–151

Contents lists available at ScienceDirect

Toxicology Letters

journal homepage: www.elsevier.com/locate/toxlet

# Statistical means to enhance the comparability of data within a pooled analysis of individual data in neurobehavioral toxicology

Monika Meyer-Baron[a,*], Michael Schäper[a], Guido Knapp[b], Roberto Lucchini[c], Elisa Albini[c], Rita Bast-Pettersen[d], Shuchang He[e], Hong Yuan[e], Qiao Niu[f], Xian-Liang Wang[g], Yong-Jian Yang[h], Anders Iregren[i], Bengt Sjögren[j], Morten Blond[k], Peter Laursen[l], Bo Netterstrom[m], Donna Mergler[n], Rosemarie Bowler[o], Christoph van Thriel[a]

[a] *IfADo – Leibniz Research Centre for Working Environment and Human Factors, Ardeystr. 67, Dortmund, Germany*
[b] *TU Dortmund University, Department of Statistics, Dortmund, Germany*
[c] *Occupational Health and Industrial Hygiene, University of Brescia, Italy*
[d] *National Institute of Occupational Health, Oslo, Norway*
[e] *Department of Psychology, Peking University, China*
[f] *Department of Occupational Health, Shanxi Medical University, Taiyuan, China*
[g] *State Key Laboratory of Environmental Criteria and Risk Assessment, Chinese Research Academy of Environmental Sciences, Beijing, China*
[h] *School of Public Health, Anhui Medical University, Hefei, China*
[i] *Toxicological Risk Assessment, Swedish Work Environment Authority, Stockholm, Sweden*
[j] *Work Environment Toxicology, Institute of Environmental Medicine, Karolinska Institutet, Stockholm, Sweden*
[k] *Clinic of Occupational Medicine, Nykoebing, Denmark*
[l] *Danish Working Environment Authority, Copenhagen, Denmark*
[m] *Clinic of Occupational Medicine, Copenhagen, Denmark*
[n] *University of Quebec, Montreal, Canada*
[o] *San Francisco State University, San Francisco, USA*

ARTICLE INFO

*Article history:*
Received 28 April 2011
Received in revised form 29 June 2011
Accepted 1 July 2011
Available online 7 July 2011

*Keywords:*
Systematic review
Meta-analysis
Manganese
Cross-cultural comparison
Confounder

ABSTRACT

Meta-analyses of individual participant data (IPD) provide important contributions to toxicological risk assessments. However, comparability of individual data cannot be taken for granted when information from different studies has to be summarized. By means of statistical standardization approaches the comparability of data might be increased. An analysis of individual data on the neurobehavioral impact of manganese (Mn) exemplifies challenges and effects of a multilevel statistical procedure.

Confounding from individual-level and study-level covariates was shown by analyses of variance, but could be reduced by linear regressions and z-normalization using data of the respective control groups. Fixed models that were used to estimate the impact of the neurotoxic exposure, provided evidence that the employed procedures, especially the z-normalization, effectively reduced variance that was unrelated to the neurotoxic exposure. Even after this statistical treatment the fixed effect models revealed differences among studies that did not seem to be exhaustively explicable by concentration differences obvious from the Mn biomarker at hand.

IPD studies using confounded endpoints as effects markers can be reasonably summarized when appropriate statistical operations are employed. For the data at hand the proposed normalization allowed new insights into exposure–effect relationships, in general it appears appropriate to investigate the effect of the independent variable more closely.



## 1. Introduction

Meta-analyses of aggregated data (AD) provide important contributions to risk assessment, but the analyses of dose–response relationships, person-specific risk factors, and covariates often remain dissatisfying (Meyer-Baron et al., 2008, 2009). Meta-analyses of IPD (individual patient data) are considered as having

*Abbreviations:* AD, aggregated data; IPD, individual patient/participant data; Mn, manganese; MnB, manganese in blood; SRT, simple reaction task; SPES, Swedish performance evaluation system.

* Corresponding author. Tel.: +49 231 1084 416; fax: +49 231 1084 308.
*E-mail address:* meyerbaron@ifado.de (M. Meyer-Baron).

0378-4274/$ – see front matter © 2011 Elsevier Ireland Ltd. All rights reserved.
doi:10.1016/j.toxlet.2011.07.003

distinct advantages in this respect; the evaluation of the validity of research findings across independent samples, measures, and designs has also been emphasized (Curran and Hussong, 2009; Hofer and Piccinin, 2009). These features appear as assets when the analysis goes beyond risk identification and tackles dose–response relationships. However, when individual data from different studies have to be summarized, the comparability of data becomes an issue. Study-level covariates, i.e. differences among the incorporated studies may confound the results in addition to individual-level covariates, i.e. inter-individual differences within a particular study.

In the study at hand, for example neurobehavioral data from North America, Europe, Africa and China were supposed to be summarized: the neurobehavioral impact of manganese was to be investigated since behavior is regarded as an early indicator of neurotoxicity (Lucchini et al., 2005, 2000; Rohlman et al., 2008). The dependent variables were performance scores obtained by neuropsychological tests. The performance is supposed to mirror a bundle of physiological and psychological processes influenced by biological, educational and motivational aspects. It was apparent that cross-cultural comparability could not be taken for granted and evidence for the cross-cultural diversity of performance scores had indeed been provided (Anger et al., 1993; Chung et al., 2003; Nell et al., 1993). Also the inclusion of differently measured individual-level covariates provided obstacles, for example when the pre-morbid intellectual capacity had to be considered. The categorical information about education years in China could not straightforwardly be compared to a measure of verbal intelligence from Sweden, although both variables claimed to reflect the pre-morbid intelligence. Differences might be surmised even within one country, when for example rural and urban samples are to be compared.

Despite efforts on investigating and modeling of the heterogeneity of data (Mathew and Nordstrom, 2010; Tudur Smith et al., 2005) little attention has been paid to these problems by most of the IPD studies in the medical field. As Simmonds et al. (2005) pointed out, not even the use of a random effects model incorporating potential heterogeneity of the endpoints was common. The typical way of analyzing the influence of covariates was the creating of subgroups with respect to a single covariate. More elaborated studies used random effects models for effect estimates (e.g. Mauer et al., 1999), sometimes differences between expected and observed values (Yusuf et al., 1985) or proportion scores were calculated (Ferrari et al., 2001); sometimes studies summarized differently measured covariates (e.g. Pocock et al., 2001).

However, each of the single approaches appears insufficient, if the summary aims at (a) using the individual data to estimate the effect across different studies, (b) considering the differently measured covariates, and (c) relating the individual data to exposure measures. Even if an effect size could be appropriately estimated by a random effects model on the basis of data from different cultures, the differences among the studies remain a “random factor” in a statistical computation and the individual data cannot be related to a common exposure measure.

We will introduce a comprehensive approach that comprises (a) a study-wise adjustment for individual-level covariates, (b) a z-normalization taking account of study-level covariates, and (c) fixed effect models to estimate the exposure-related effects within our sample of studies.

The approach is universally applicable to data that differs as a function of study-level covariates enhancing the variance unrelated to independent variables like exposure or treatment for example. Since active workers were under scrutiny in our analysis, “IPD” abbreviates “individual participant data” as proposed before (Cooper and Patall, 2009).

## 2. Materials

The same sample as in our previous AD meta-analysis on Mn (Meyer-Baron et al., 2009) was considered eligible. Studies had been excluded because of non-random samples, sporadically employed neuropsychological tests, a lack of information on internal exposure, or re-examination of participants. Each of these shortcomings precluded the studies also from the pooled analysis. Since our final computations started in February 2009, the studies by Chang et al. (2009) and Cowan et al. (2009) were not considered.

The corresponding authors of the eligible studies were informed about the objectives of the study by a cover letter. Where agreement was obtained, a contract about the confidential use of the data was signed. The supplied anonymous data were checked for congruence with the published information and plausibility compared to data from other studies. Equivocality was resolved by communication with the researchers.

The raw data of the component studies were re-named in a common way and a master data set created. Not all performance tests were available at study-level; they were analyzed in the respective subsample. In one study (Wang et al., 2006) some of the MnB values appeared not to be reliable. In agreement with the principle investigator the detection limit was used for 7 exposed and 12 unexposed subjects.

## 3. Methods

Our methodological approach comprised the check of the data for certain influences first and the application of the statistical operations thereafter. Both steps will be described in the following; the statistical operations are depicted schematically in Fig. 1.

1. Relations among performance scores were explored by correlations calculated separately for cognitive and motor performance tests. This was done within each study, because the administered tests differed among the studies.
2. Confounding from individual-level covariates was explored by stepwise regressions within each study. Influences from age, pre-morbid intelligence, alcohol intake, and smoking habits were analyzed, because (a) information on these covariates were available from almost all studies, and (b) the influences of these covariates on neuropsychological test performances are substantial, as shown before (Cervilla et al., 2000; Glass et al., 2009; Hedden and Gabrieli, 2004; Kalmijn et al., 2002; Richards et al., 2003; Whalley et al., 2005).
   Separate multiple linear regressions were calculated for the scores of each neuropsychological test within each study. The most detailed format of the covariate was used, e.g. number of cigarettes instead of smoker vs. non-smoker. In this way (a) the number of covariates did not vary among tests and studies, and (b) there was a global set of confounders, but strengths of individual studies were considered. The resulting performance scores were denoted “adjusted scores”.
3. The need for considering study-level covariates was explored by ANOVAs using the factor STUDY and subsequent pair-wise comparisons. The adjusted data of the reference groups were used because (a) the differences should reflect differences that were unrelated to the obvious exposure differences among the exposed groups (see Table 1), and (b) it should be avoided that differences were attributed to cross-cultural differences although they were explicable by inter-individual differences. Because of influences from unknown study-level covariates, indicated by a significant main effect of STUDY and significant contrasts in the pair-wise comparison, the adjusted scores of the participants were “z-normalized”. We chose this term to point out that not the mean of the total sample was used but of the reference group of the respective study. The mean performance score of the reference group was subtracted from the individual performance score and the difference divided by the SD of the reference group. The data of the reference subjects substituted the normative data that were not available for each test and country. The scores were denoted “z-normalized scores”.
4. Fixed effect ANOVAs were used to estimate exposure-related differences across the included studies (STUDY) and between the exposed and control groups (GROUP). All effects were modeled as fixed effects; variances were allowed to differ across the individual studies.

The main effect GROUP was supposed to largely reflect a general effect of the neurotoxic exposure (exposed workers vs. controls). However, additional exposure-related differences might be introduced by the particular exposure conditions of the included studies. Therefore, the computation of the main effect STUDY was restricted to the exposed participants of the different studies. The interaction GROUP × STUDY also reflected exposure-related differences, namely the modulation of the general exposure effect due to concentration differences among the studies. More precisely, a stronger effect of the factor GROUP would be expected in studies investigating significantly higher exposed workers. Statistically, this would be reflected by a significant interaction GROUP × STUDY.

In order to estimate the effects of the three preceding steps on exposure-related differences, the computations were run for the different types of data generated in the stepwise transformation approach (raw, adjusted, and z-normalized performance scores).