HEALTH RISK ASSESSMENT GUIDANCE FOR METALS

FACT SHEET

# HERAG 06

## QUALITY SCREENING PROCEDURES FOR HEALTH EFFECTS LITERATURE

**August 2007**
**Check you have the most recent fact sheet by visiting www.metalsriskassessment.org**

EBRC

EUROFER European Confederation of Iron and Steel Industries

EuroMetaux

ICMM International Council on Mining & Metals

**Table of contents**

1. Introduction ..... 3
2. Quality screening procedures for health effects study reports and publications ..... 3
2.1. Animal experimental investigations ..... 3
2.2. Human data ..... 5
2.3. Genotoxicity/mutagenicity data ..... 5
3. Previous experience in EU risk assessments ..... 5
3.1. Copper ..... 5
3.2. Zinc ..... 6
3.3. Lead ..... 6
3.4. Cadmium ..... 7
3.5. Nickel ..... 8
4. Conclusions and recommendations ..... 8
5. References and Abbreviations ..... 10
Appendix 1. Examples of quality scoring systems applied in previous EU Risk Assessments ..... 12
A 1.1. Quality criteria for mutagenicity studies ..... 12
A 1.2. Rating system for the quality of deficiency studies (Source: VRA Copper) ..... 12
A 1.3. Rating system for human studies (Source: VRA Lead) ..... 15
A 1.4. Examples of a checklist used in the Cd-RA to evaluate studies on human health ..... 23
A 1.5. Toxicological study reliability code documentation (Source: NiPERA) ..... 28

## 1. Introduction

The aim of this fact sheet is to provide a common outline of quality screening procedures for published and unpublished reports on human health effects, by which the quality/reliability and relevance of reports can be assessed.

Whereas this may intrinsically not appear to be a metal-specific issue, the experience of data-rich substances (such as lead) shows that for a wide range of studies, quality and relevance scoring systems had to be developed.

As a consequence, such procedures may be anticipated to differ in detail and extent between (i) animal testing reports, and (ii) human epidemiological studies, for example. It was therefore decided to establish this fact sheet as a common position of the metal industries previously involved in risk assessment exercises.

Prior to drafting this fact sheet, all key metal industries were approached by a questionnaire, and prompted to provide (where available) a set of criteria that they may have previously applied to judge the relevance and quality/reliability of health effect literature in their metal risk assessment.

## 2. Quality screening procedures for health effects study reports and publications

It is common practice in an EU regulatory context to validate animal experimental studies for chemical substances according to the scheme by Klimisch et al. (1997). In brief, any source document is screened according to a set of criteria, and finally assigned a reliability index (1-4), as outlined further below. The use of the Klimisch reliability codes allows the reader to easily locate and initially focus on the most reliable studies first. In contrast, studies which fail to meet essential reliability criteria are set aside from the beginning.

### 2.1. Animal experimental investigations

(I) The Klimisch et al (1997) scoring scheme

The original scoring system developed by Klimisch and co-workers within the context of the German government's national programme on Existing Chemicals („BUA") takes into consideration three aspects, which it defines as follows:

**Reliability:** evaluating the inherent quality of a test report or publication relating to preferably standardised methodology and the way the experimental procedure and results are described to give evidence of the clarity and plausibility of the findings.

**Relevance:** covering the extent to which data and tests are appropriate for a particular hazard identification or risk characterisation.

**Adequacy:** defining the usefulness of data for hazard/risk assessment purposes. When there is more than one study for each SIDS element, the greatest weight is attached to the study that is the most reliable and relevant. Robust study summaries are prepared for the highest quality or "key" studies.

In practice, the Klimisch scoring system categorises the reliability of a study as follows:

**1 = reliable without restrictions**:
"Studies or data ... generated according to generally valid and/or internationally accepted testing guidelines (preferably performed according to GLP) or in which the test parameters documented are based on a specific (national) testing guideline ... or in which all parameters described are closely related/comparable to a guideline method."

**2 = reliable with restrictions:**
"Studies or data ... (mostly not performed according to GLP), in which the test parameters documented do not totally comply with the specific testing guideline, but are sufficient to accept the data or in which investigations are described which cannot be subsumed under a testing guideline, but which are nevertheless well documented and scientifically acceptable."

**3 = not reliable**:
"Studies or data...in which there were interferences between the measuring system and the test substance or in which organisms/test systems were used which are not relevant in relation to the exposure (e.g., non physiological pathways of application) or which were carried out or generated according to a method which is not acceptable, the documentation of which is not sufficient for assessment and which is not convincing for an expert judgment."

**4 = not assignable**:
"Studies or data....which do not give sufficient experimental details and which are only listed in short abstracts or secondary literature (books, reviews, etc.)."

It is noted that the OECD in their Existing Chemicals Programme have developed a "Manual for Investigation of HPV Chemicals", and within this a specific guidance for scoring the data for reliability, relevance and adequacy, where the original Klimisch scoring system is fully implemented. We also note that in addition to the initial scoring process, the OECD also foresees a "weight-of-evidence"-analysis.

(II) Subsequent adaptation (TNO)

The Klimisch scoring system has subsequently been applied in adapted and modified form (TNO), where it is cited as follows: Reliability index and usefulness of information (within the framework of Council Reg. 793/93/CEE and ComReg. 1488/94 (Klimisch H.J., Andreae M., Tillmann U. (1996), adapted by TNO/RIVM (1997) and modified. The outline of this TNO adapted procedure, which deviates somewhat from the original Klimisch approach, is summarised below:

| **reliability index** | **result** | **description of reliability** |
|---|---|---|
| 1 | reliable without restrictions | the method and description are in accordance with test guidelines |
| 2 | reliable with restrictions | the method and/or description are less in accordance with test guidelines |
| 3 | not reliable | the method and/or description are not in accordance with test guidelines |
| 4 | not assignable | the original data are not available |
| | | |
| 1. example | - complete test report available: GLP, Annex V, OECD, EU etc…<br>- publications are not included | |
| 2. example | - validity of data cannot be fully established<br>- some modifications or omissions in method and description<br>- acceptable publication (e.g. according to EU- or OECD guidelines) | |
| 3. example | - method unknown and/or critical pieces of information are not available (e.g. identity of the substance)<br>- documentation not sufficient for unequivocal assessment<br>- do not meet important criteria of today's standard test methods | |
| 4. example | - only abstract available<br>- secondary literature (reviews, tables, etc…) | |

### 2.2. Human data

The assessment of the quality and relevance of human data is a complex matter that has apparently been dealt with at considerably varying levels of detail in previous EU risk assessments. For this reason, any such previous experience that is known to the authors of this document is summarised briefly on a metal-by-metal basis in chapter 3 below and relevant examples of quality scoring templates are presented in the appendix to this fact sheet.

### 2.3. Genotoxicity/mutagenicity data

The use of any such criteria was neither documented in any previous EU risk assessment for metals, nor made available prior to finalisation of this fact sheet. However, metal specific issues of mutagenicity test were extensively discussed at the HERAG Workshop on Genotoxicity/Mutagenicity and Carcinogenicity of Metals (see separate fact sheets on Mutagenicity and Carcinogenicity). This fact sheet notes that IWGT (International Workshops on Genotoxicity Testing) has recommendations (e.g. in vitro micronucleus), but these criteria have not yet been "centralised" or published (see also Appendix A 1.1 to this fact sheet).

## 3. Previous experience in EU risk assessments

The following subchapters in this section summarise the quality screening procedures as they were either followed or documented in the respective risk assessments. The authors of this fact sheet do not suggest that in any particular case such procedures were not followed – however, not in all cases is the documentation complete and transparent.

The following examples (all extracted from experience with previous risk assessments) illustrate that the diversity between nature and extent of data sets for various metals implies that there likely can not be one common approach for all metals, and that a case-by-case approach needs to be developed.

This fact sheet may help to select the most suitable approach, or at least provide guidance on relevant aspects which need to be considered in such an exercise.

### 3.1. Copper

(1) animal experimental data:

In the voluntary Copper risk assessment, the entire set of animal experimental data were subjected to a screening procedure according to the Klimisch grading system (the body of the report cites Klimisch et al., 1997), the results of which were documented in detail in the risk assessment report for each animal study. The methodology paper for the risk assessment further describes the procedure as follows:

i) The requirements for standard test methods (e.g. by EC, OECD or EPA) and GLP principles will be regarded as a reference when evaluating test data for the end-points under consideration.

ii) The benchmark for toxicity testing will be conform with Annex V to Directive 67/548/EEC

iii) Animal studies that broadly conform to Annex V standards and endpoints will be included. Deviations from Annex V protocols will be discussed in the text. Informative studies, which do not follow Annex V protocols or are not specifically directed towards Annex V endpoints, will be included on individual merit where they add significantly to the discussion.

iv) All GLP tests reports should be available in full for their independent evaluation by the research team.

v) General evaluation criteria for *in vitro* studies will include how well the test substance is characterised, the range of exposures used, whether appropriate positive and negative controls have been included, adequate number of replicates used and how well the study is reported.

(2) human data

Toxicity: the methodology document lists some general criteria, but the application of such is not documented in the body of the risk assessment report. However, this is of marginal relevance since the major conclusions of the RA on human health effects due to toxicity do not in fact rely essentially on human data.

Deficiency: it is explicitly noted here that in the chapter on deficiency effects, a separate "utility rating system" for study quality of deficiency studies was developed (see A 1.2).

### 3.2. Zinc

Despite that one may assume that this was done, it is not explicitly documented in the RAR on zinc and zinc compounds whether human health effects data for zinc were assessed for study quality, and also not by which *a priori* strategy. Health effects data were summarised in narrative review fashion with the relative strengths and weaknesses of each study being noted. This approach would appear to be less rigorous than that recently applied for substances such as lead (see below), and is more common for substances and/or health effects that are less than "data rich". Although some element of subjectivity can be evident in narrative review approaches, more detailed, standardised review strategies can successfully identify the higher quality studies in the literature.

### 3.3. Lead

(i) animal experimental data

Animal experimental data were incorporated into the RAR merely for the purpose of classification and labelling, i.e., on acute (oral, dermal, inhalation) toxicity, skin and eye irritation, and skin sensitising properties. All such studies were subjected to a scoring system according to Klimisch et al (1997).

(ii) human data

The risk assessment on lead and lead compounds is exclusively based on health effects data in humans. Such health effects data were collected for multiple endpoints of potential concern for the repeated dose toxicity of lead. Health endpoints reviewed included: haematological, renal, carcinogenic, genotoxic, cardiovascular, reproductive, developmental and neurobehavioral effects. Given that lead and lead compounds have been extensively studied, multiple publications had evaluated potential impacts for each of the above endpoints. In some instances, more than 100 studies had been conducted of a single toxicological endpoint and divergent estimates offered of dose response and/or effects. In order to assist in evaluation of this extensive scientific literature, detailed study quality criteria were developed for toxicological endpoints of principle concern. These study evaluation criteria were specific to individual health endpoints and the study designs that have evolved over the years to evaluate lead health effects.

For critical health endpoints evaluated in Repeated Dose Toxicity, each study was scored for the extent to which key elements of study design met rigorous quality criteria. Although these criteria were different for each endpoint, criteria typically included study adequacy in terms of size and power to investigate the health effect of concern, the extent to which exposed and control populations were

matched, care exercised in evaluating current lead exposure and lead exposure history, whether relevant co-exposures had been monitored, ascertainment and correction for important confounders, and quality of the techniques used to assess health endpoints. Each study was assigned a numerical score that increased as a function of the extent to which specific key elements of study design had been executed and described. Since the highest numerical value that a study could attain varied with each health endpoint and type of study, study quality scores were converted to percentiles that reflect the overall extent to which each study met the quality criteria. The perfect study would thus have an overall percentile score of 100%. The percentile quality scores were included in the summary tables presented at the end of each health effects endpoint section and were used to support the "weight of evidence" conclusions reached in the derivation of NOAEL's and the validity of effect sizes and dose response functions suggested by narrative reviews and meta-analyses. Examples for such a grading systems are given in the Appendix (A 1.3) to this fact sheet.

### 3.4. Cadmium

The aim was to base the risk assessment on cadmium metal and cadmium oxide on health effects data in humans, as extensive scientific literature was already available regarding most of the endpoints of potential concern. Animal experimental data were only included in the RAR when reliable human data were not available for a specific endpoint or for further elaboration on mechanisms of toxicity. Because of the amount of data, efforts were made to develop a methodology in order to facilitate the search for, and evaluation of, relevant publications.

Data search:

The dataset made available by industry was considered as a starting-point, and supplemented by four existing literature reviews with international credit in order to identify the publications of relevance (IARC (1992, 1993); ATSDR (1993, 1999); Friberg et al. (1985, 1986), IPCS (1992)).

It was not satisfactory to simply carry over the conclusions of these reviews for the purposes of Cd metal/CdO hazard assessment, as

a) the data included in these reviews were not systematically evaluated, and
b) most of these reviews did not take speciation into account, leaving uncertainties as to the effects specifically associated with the two Cadmium compounds to be assessed (Cadmium metal and Cadmium oxide).

Consequently, the original reports cited in these reviews were consulted (including their reference lists), and a search was carried out into databases such as Medline, Toxline, TOMES etc., using keywords referring to the health endpoint of concern.

Data quality/reliability:

(i) animal data: experimental studies using cadmium oxide or cadmium metal and critical studies using cadmium compounds were described in general. All these studies were subjected to a scoring system in accordance with Klimisch et al. (1997).

(ii) human data: all the studies identified were evaluated using checklists relating to population, exposure, endpoints, biases and confounders (see example checklists in Appendix A 1.4). These checklists were drawn up by type of study design. The use of such checklists enables the strengths and weaknesses of a study, its potential information gaps (e.g. on the cadmium compound in presence), and possible confounding or modifying factors to be analysed. No score was given to a particular study. This evaluation process made it possible to identify certain sources of heterogeneity between the results of all available studies for a specific endpoint and to take these into account in the overall assessment of the weight of evidence.

### 3.5. Nickel

NiPERA has provided examples of data quality screening protocols (see Appendix A 1.5 to this fact sheet). However, in assessing the approach used by the Danish rapporteur in the EU Risk Assessments (RAs) of nickel and nickel compounds, no established criteria for evaluating study reliability were apparently applied in the human health component of these RAs. However, the Danish rapporteur explicitly required data quality assessments in the environmental component of the nickel/nickel compound RAs.

In the Human Health component of the RAs, the Danish rapporteur performed a subjective case-by-case assessment of study quality, discussing both good and bad studies in the text of the RA. In the opinion of the nickel industry, it would have been preferable had the establishment of objective study quality criteria been applied, but the need to do so was less obvious in the human health assessments given that the reference studies were well known and generally agreed upon between the rapporteur and industry. The area that was most prone to disagreement between industry and the Danish rapporteur was the interpretation of the reference study results and the extrapolation of those data to actual calculations of risk.

## 4. Conclusions and recommendations

This fact sheet summarises a retrospective analysis of the quality screening procedures as they were either followed or documented in previous metals risk assessments, and as they were made available to the authors.

Given the variance in extent and the scope of human health data between metals, it is unlikely that a “standard” one-for-all approach can be taken, and that instead a case-by-case approach may be more appropriate.

This examples presented in the appendix to this fact sheet may help to select the most suitable approach, or at least provide guidance on relevant aspects which need to be considered in such an exercise.

In the discussion of these aspects within the HERAG project community, the following general recommendations and/or suggestions were made:

- The use of suitable biomarkers is considered to enhance the credibility of a particular report/study, and of an entire data base where used consistently. It is therefore recommended to emphasise this aspect (examples: lead and cadmium).

- The relevance of confounders is significant, both on the production and downstream user level when dealing with epidemiological data for workers, as well as in consumer / general population exposure studies.

- Background data: “natural” background levels exist for all metals; the relevance of reported effect levels and exposure monitoring data should be considered against these, where possible.

For animal toxicity studies, the Klimisch grading system has been most consistently and widely applied. It was found to be easily workable, and corresponds well to the requirements of the IUCLID data base for documenting the influence of methodology and quality of a study on the results. Ranking individual experimental data by this scheme has facilitated transparency in documenting how health end-points for risk characterisation were selected based upon a weight-of-evidence approach.

For human studies, no single approach was found to be clearly superior to all others. In some cases, an end-point by end-point scheme had to be developed. However, in doing so, it was also recognised as important to consider the level of detail required for some of the approaches listed in the

appendices to this fact sheet. If too much detail is required, the screening criteria may become so burdensome that it will be unlikely to be readily adopted. Therefore, the final procedure should be simple enough to not require excessive resources to complete while providing sufficient clarity regarding the quality of the study.

In addition to the above, the following metal-specific issues were considered by the HERAG project group worthy of mentioning:

- Despite as being discussed for their relevance in the setting of $MOS_{ref}$ values, aspects such as genetic sensitivity, nutritional status for essential trace elements and possible relevance of background exposure/intake for non-essential elements may warrant consideration.

- For epidemiological studies, the particular relevance of Arsenic as a confounder in many primary metal producing industries, and in more general terms, the adequate assessment of co-exposures were stressed as being important.

- Also for epidemiological studies, the quality of the retrospective exposure assessment is a decisive factor (see also separate HERAG fact sheets on Carcinogenicity); several approaches in the past have been developed for retrospective exposure assessments of metals which are intrinsically prone to a high level of error.

- Chemical speciation: the aspect of speciation should be explicitly included in the evaluation of any study report or publication. It should include the physicochemical characteristics of the test- or study substance in relation to the metal species for which the risk assessment report is written. It is an important element when discussing relevance and adequacy of the data.

As a consequence from the quality scoring procedures applied, the summary of health effects data should reflect this previous work effort by (i) clearly identifying the data collection processes and the selection criteria for each endpoint for reasons of transparency, and (ii) a discussion of the homogeneity or heterogeneity of the data. If sufficient quantitative data are available, a meta-analysis should be applied (example: VRA Lead).

## 5. References and Abbreviations

References

ATSDR (1993): Toxicological Profile for Cadmium. ATSDR (Agency for Toxic Substances and Disease Registry). Atlanta, Georgia 30333 U.S. Department of Health & Human Services, Public Health Service (1993)

ATSDR (1999): Toxicological Profile for Cadmium (Update). ATSDR (Agency for Toxic Substances and Disease Registry). Atlanta, Georgia 30333 U.S. Department of Health & Human Services, Public Health Service (1999). Internet: http://www.atsdr.cdc.gov/toxprofiles/tp5.html [as of 2005-06-23]

Friberg, L. et al. (1985): Cadmium and Health: a toxicological and epidemiological appraisal. Volume I: Exposure, Dose and Metabolism. Edited by L Friberg, CG Elinder, T Kjellström, and GF Nordberg. CRC Press, Boca Raton, Florida (1985).

Friberg, L. et al. (1986): Cadmium and Health: A Toxicological and Epidemiological Appraisal. Volume II: Effects and Response. Edited by L Friberg, CG Elinder, T Kjellström, and GF Nordberg. CRC Press. Boca Raton, Florida (1986).

IARC (1992): IARC (International Agency for Research on Cancer). Cadmium in the Human Environment: Toxicity and Carcinogenicity. (1992)

IARC (1993): IARC (International Agency for Research on Cancer). Cadmium and Cadmium Compounds. In:Volume 58: Beryllium, Cadmium, Mercury and Exposures in the Glass Manufacturing Industry. Edited by IARC LYON. IARC. United Kingdom, 119-237. (1993)

IPCS (1992): IPCS, International Programme on Chemical Safety (1992). Cadmium. Environmental Health Criteria 135. World Health Organization, Geneva.

IPCS (2002): Principles and Methods for the Assessment of Risk from Essential Trace Elements, IPCS/WHO, Geneva, Series Environmental Health Criteria, No. 228, ISBN 92-4-157228-0

Klimisch H.J. et al. (1997): A systematic approach for evaluating the quality of experimental and ecotoxicological data, Reg.Tox. and Pharm. 25, 1-5 (1997)

Plunkett, L. (2004): Personal communication, March, 2004.

Abbreviations

| | |
|---|---|
| DK | Denmark |
| DRC | Dose-response curve |
| EC | European Community |
| ECG | electrocardiogram |
| EPA | Environmental Protection Agency (USA) |
| ETE(s) | essential trace element(s) |
| FAB (criteria) | French-American-British to group for instance myelodysplastic syndromes, refractory anaemia etc |

| | |
|---|---|
| GLP | Good Laboratory Practice |
| HPV | High production volume chemical |
| ICA | International Copper Association |
| ICD | International Classification of Diseases: A system of categories to which morbid entries are assigned according to established criteria. Included is the entire range of conditions in a manageable number of categories, grouped to facilitate mortality reporting. |
| ICD-O | International Classification of Disease for Oncology |
| NiPERA | Nickel Producers Environmental Research Association |
| NOAEL | no observed adverse effect level |
| OECD | Organisation for Economic Co-operation and Development |
| RA(R) | Risk Assessment (Report) |
| SES | Socioeconomic status |
| SIDS | Screening information data set |
| SLE | systemic lupus erythematosus classification/criteria |
| SOD | superoxide dismutase |
| TGD | Technical Guidance Document |

## Appendix 1. Examples of quality scoring systems applied in previous EU Risk Assessments

### A 1.1. Quality criteria for mutagenicity studies

In the discussions at the Scientific Review Panel level, a suggestion was made to also include an example of criteria for genotoxicity/mutagenicity studies, which was previously missing in this fact sheet.

The establishment of general quality criteria for mutagenicity tests was not a key topic of the HERAG project. Nevertheless, it is noted that the use of any such criteria was neither documented in any previous EU risk assessment for metals, nor made available prior to finalisation of this fact sheet. However, metal specific issues of mutagenicity test were extensively discussed at the HERAG Workshop on Genotoxicity/Mutagenicity and Carcinogenicity of Metals (see separate fact sheets on Mutagenicity and Carcinogenicity).

General quality criteria for such test were only briefly discussed at this workshop. While there are basic do's and don'ts that run across the different tests, it is difficult to come up with general guidelines. The fact sheet on Mutagenicity notes some of the test protocols for which IWGT (International Workshops on Genotoxicity Testing) has recommendations (e.g. in vitro micronucleus). However, these criteria have not yet been "centralised" and published, partly due to the fact that guidelines for the different mutagenicity tests are constantly evolving. In addition, classification and labelling decisions where such test play an important role, involve detailed evaluation of study data by experts familiar with the current state of the art for each test, however not following a given set of fixed quality criteria.

### A 1.2. Rating system for the quality of deficiency studies (Source: VRA Copper)

Study Quality Criteria

The TGD provides guidance for evaluating the quality of toxicity studies; a set of evaluative criteria has been developed by Klimisch et al. (1997) to assess and rate these studies using a “reliability index” scale of 1-4. This reliability index establishes the adequacy of the study for assessment of chemical toxicity and is based on the degree of concordance of the study with toxicology test guidelines established by regulatory bodies such as OECD and EU Annex V. The criteria include (1) the degree of availability of data for independent evaluation (e.g., a complete test report from the laboratory versus an acceptable publication versus an abstract or secondary source reference); (2) the degree of completeness of reporting; and (3) adherence to test guidelines for study variables such as chemical form and purity; animal housing conditions; study design and methodology; data validation; statistical analysis; and presentation of relevant data.

Impairment of health associated with trace element deficiency is considered to be a serious public health problem among nutritionists, public health specialists, and molecular biologists but until recently has been of low concern among toxicologists and chemical risk assessors. The TGD and Klimisch et al. (1997) criteria were designed for xenobiotic chemicals with no requisite physiological functions and whose presence in the body at any dose indicate a potential for toxicity. These criteria were not designed to assess the potential adverse consequences of ETEs, for which physiological, structural and/or functional impairment due to either excessive or deficient exposures is of concern. Regulatory test guidelines for conducting studies to characterize the hazards and dose-response of trace element deficiency for the purpose of risk characterization have not yet been developed; the need to develop such test guidelines has been discussed in IPCS (2002). Further, deficiency studies have typically been designed to identify the adverse health consequences of severe deficiency and only one dose has been tested, limiting the use of these findings for dose-response assessment. Nonetheless, there is guidance for design and conduct of nutrient deficiency studies developed by various nutrition boards and societies that follows sound scientific principles and good laboratory practices (see, for example, the Nutrition Society [Europe], American Society for Nutritional Sciences; American Society for Clinical Nutrition). These studies have generated reliable data on hazards, with some utility for dose-response assessment.

As part of an ongoing dose-response assessment project for the International Copper Association (ICA), a set of evaluative criteria for rating deficiency studies for quality, utility in risk assessment, and degree of concordance with toxicity test guidelines has been developed. These criteria, with some modifications, have been used to rate the deficiency studies reported in this Appendix and are presented in Table 1 (next page). As with all evaluative criteria, including those of Klimisch et al. (1997), study quality on a case-by-case basis is also assessed using best professional judgment.

Table 1: Utility Rating System (Study Quality) for Deficiency Studies in Humans and Laboratory Animals[1]

| 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|
| Use of multiple doses in intact humans or animals; dose-response data<br><br>Identification of multiple outcomes;<br><br>Good design, conduct and reporting based on sound scientific principles and laboratory practices<br><br>Outcomes measured included functional and/or structural effects at the level of the organism, organ, tissue, or cell.<br><br>Best professional judgment. | One or two doses from intact animals or humans<br><br>Identification of multiple outcomes, usually in targeted (one) organ or system.<br>Temporal changes in outcomes measured.<br><br>Good design, conduct and reporting; based on sound scientific principles and laboratory practices;<br>Potentially useful or limited data for dose-response.<br><br>Outcomes measured may have included functional/structural effects as well as cellular/intracellular effects<br><br>Best professional judgment. | Single dose or clinical study/case report with some dose information<br><br>Studies only showing interaction of two or more independent variables<br><br>Fairly good reporting<br><br>Tracer or pharmacokinetic studies<br>Information on body burden, kinetics/dynamics<br>Mechanistic or intracellular effects only.<br>Best professional judgment | No dose information or physiological end points only<br><br>Review<br><br>Best professional judgment | No utility<br>No dose information by relevant route(s) of exposure |

[1] Adapted from criteria developed for rating the utility and quality of studies of excess or deficiency of essential trace elements, described by Plunkett (2004), and presented at the U.S. Society of Toxicology Meeting, Baltimore, MD, March, 2004.

Severity-of-Effect Criteria

In addition to ratings for study quality and utility, observed effects due to chronic deficiency were also rated for severity. A severity scale developed for the ICA dose-response assessment project was used, with some modifications. The severity ratings ranged from 1D (deficiency) to 4D. In general a rating of 1D represents physiological changes that are considered to be homeostatic responses to fluctuations in Copper bioavailability and are interpreted as being adaptive rather than adverse, whereas a rating of 4D is indicative of severe clinical outcomes or irreversible changes in structure or function associated with clinical impairment. Ratings of 2D and 3D represent increasing severity of effects that fall between these two ends. The types of end points assigned to each severity rating is presented in more detail in Table 2.

Table 2 Severity Scoring For Deficiency Studies in Humans and Laboratory Animals2

| Severity Score | Types of End Points |
|---|---|
| 1 D | Changes in Cu body burden; tissue stores, metallothionein levels, Cu excretion; considered to be adaptive and within the range of controlled homeostasis |
| 2 D | Decrease/loss of plasma/red cell Copper-dependent enzyme function (e.g., ceruloplasmin, diamineoxidases, SOD) without histopathology; changes in blood/liver lipids (e.g., cholesterol, triglycerides); mild-to-moderate body weight decreases; changes in red blood cells (# or function) |
| 3 D | Severe body weight decreases (>20%); changes in organ weight(s) or enzymes indicative of organ dysfunction (e.g., liver), plasma glucose/insulin, heart rate/blood pressure/ECG, white blood cell (#, function)/other indices of impaired immune function; anaemia, haemolysis; inflammation; histopathology or changes in ultra-structure; neuromuscular/neurobehavioral changes; alterations in hormone levels |
| 4 D | Mortality; gross pathology (e.g., bone deformities); changes in reproductive function indices (e.g. failure to reproduce; teratology) |

[2] Adapted from a scoring system developed to rate the severity of effects observed in studies of excess or deficiency of essential trace element, Plunkett (2004).

### A 1.3. Rating system for human studies (Source: VRA Lead)

Human health evaluation tables – Lead Risk Assessment:

A "score" (from 1 to 3) is given for each parameter according to the following table. The overall "reliability factor" is established by simple addition of the separate scores: low: from 14 to 21, medium: from 21 to 35; high: from 35 to 42.

STUDY ANALYSIS - KIDNEY – OCCUPATIONAL

| Source<br>Title:<br>Author:<br>Journal:<br>Study Type: prospective – longitudinal – cross-sectional<br>Predominant exposure source (if known): | | | | | | |
|---|---|---|---|---|---|---|
| | High = 3 | Medium = 2 | Low = 1 | | | |
| Cohort Definition and Size | | | | | Average Score | |
| Study population<br>-sample size and power | Sufficient power for anticipated effect | Marginal power for anticipated effect | Inadequate power | | | |
| Selection/participation bias<br>☐ Sample recruitment (exclusion/inclusion criteria)<br>☐Sample characteristics<br>☐Characteristics (and number/%) of eligible non-participants | Well described | Poorly described | Not described | | | |
| Control group<br>(If applicable) | Present, highly comparable to the exposed group | Present, comparable to the exposed group | Absent | | | |
| Exposure | | | | | Average Score | |
| Exposure measure | Blood-leads | Air, urine, bone or other measure | None | | | |
| Exposure history<br>-frequency, duration, intensity | Precisely defined | Reconstructed | Unknown | | | |
| Analytical quality control of exposure assessment | Well documented | Roughly documented | Not documented | | | |
| Endpoint Definition | | | | | Average Score | |
| Endpoint | Accepted and well defined | Equivocal significance | Uncertain significance | | | |
| | | | | | | |
| Results | Clearly reported | Reported | Not reported | | | |
| Dose response relationship | Examined | | Not examined | | | |
| Confounders/Control | | | | | Average Score | |
| List of potential confounders/covariates | Comprehensive and relevant | Less comprehensive but relevant | Inadequate | | | |
| Major Confounding Factors:<br>Age<br>Analgesics<br>Alcohol<br>Diabetes<br>Dietary habits<br>Other occupational nephrotoxins, e.g., Cadmium | | | | | | |
| Associations of potential confounders with exposure/effect | Documented for both | Documented for one only | Not documented | | | |
| Assessment of relevant co-exposures | Yes, several | Yes, one | None | | | |

| | |
|---|---|
| **OVERALL TOTAL SCORE** | |
| **Additional Comments:** | |

STUDY ANALYSIS - NEUROLOGICAL – OCCUPATIONAL

| | | | | | | |
|---|---|---|---|---|---|---|
| Source<br>Title:<br>Author:<br>Journal:<br>Study Type: prospective – longitudinal – cross-sectional<br>Predominant exposure source (if known): | | | | | | |
| | High = 3 | Medium = 2 | Low = 1 | | | |
| Cohort Definition and Size | | | | | | Average Score |
| Study purpose | Precisely defined | Defined without precision | Not defined | | | |
| Study population<br>-sample size and power | Sufficient power for anticipated effect | Marginal power for anticipated effect | Inadequate power | | | |
| Selection/participation bias<br>☐ Sample recruitment (exclusion/inclusion criteria)<br>☐ Sample characteristics<br>☐ Characteristics (and number/%) of eligible non-participants | Well described | Poorly described | Not described | | | |
| Control group<br>(If applicable) | Present, highly comparable to the exposed group | Present, comparable to the exposed group | Absent | | | |
| Exposure | | | | | | Average Score |
| Exposure measure | Serial blood-leads | Single concurrent plus additional retrospective marker, as e.g. tooth, dentine or bone Lead | Single concurrent blood Lead | | | |
| Exposure history: frequency, duration, intensity | Precisely defined | Reconstructed | Unknown | | | |
| Analytical quality control of exposure assessment | Well documented | Roughly documented | Not documented | | | |
| Examiner | | | | | | Average Score |
| Examiner | Qualified trained professional | Trained non-professional | Details not given | | | |
| If several examiners | Inter-observer reliabilities given | Inter-observer reliability monitored but not given | Inter-observer reliabilities not given | | | |
| Blindedness | Blinded | Doubtful | Not blinded | | | |
| Quality criteria: operator training, standardization of measurement method, errors inherent in equipment used | Adequate | Partial | Inadequate | | | |
| Endpoint Definition | | | | | | Average Score |
| Endpoint | Accepted and well defined | Equivocal significance | Uncertain significance | | | |
| Effect measure | Established clinical tool with good quality criteria | Less established tool with some quality criteria | Ad hoc measure without quality criteria | | | |
| Results | Clearly reported | Reported | Not reported | | | |
| Dose response relationship | Significant | Marginal | Not examined | | | |
| Confounders/Control | | | | | | Average Score |
| List of potential confounders/covariates | Comprehensive and relevant | Less comprehensive but relevant | Inadequate | | | |
| Major Confounding Factors:<br>Age<br>Sex<br>SES<br>o Pre-employment neuropsychological function<br>Alcohol consumption<br>Smoking<br>o Previous exposure history<br>o Education level<br>o Income/job quality<br>o Other Exposures | | | | | | |
| Associations of potential confounders with exposure/effect | Documented for both | Documented for one only | Not documented | | | |
| Assessment of relevant co-exposures | Yes, several | Yes, one | None | | | |
| **OVERALL TOTAL SCORE** | | | | | | |
| **Additional Comments:** | | | | | | |

STUDY ANALYSIS - CARCINOGENICITY – OCCUPATIONAL

| | | | | | | |
|---|---|---|---|---|---|---|
| Source<br>Title:<br>Author:<br>Journal:<br>Study Type: prospective – longitudinal – cross-sectional<br>Predominant exposure source (if known): | | | | | | |
| | High = 3 | Medium = 2 | Low = 1 | | | |
| Cohort Definition and Size | | | | | | Average Score |
| Study population<br>-sample size and power | Sufficient power for anticipated effect | Marginal power for anticipated effect | Inadequate power | | | |
| Selection/participation bias<br>☐ Sample recruitment (exclusion/inclusion criteria)<br>☐ Sample characteristics<br>☐ Characteristics (and number/%) of eligible non-participants | Well described | Poorly described | Not described | | | |
| Control group<br>(If applicable) | Present, highly comparable to the exposed group | Present, comparable to the exposed group | Absent | | | |
| Exposure | | | | | | Average Score |
| Exposure measure | Blood-leads | Air, urine or other measure | None | | | |
| Exposure history<br>-frequency, duration, intensity | Precisely defined | Reconstructed | Unknown | | | |
| Analytical quality control of exposure assessment | Well documented | Roughly documented | Not documented | | | |
| Examiner | | | | | | Average Score |
| Examiner | Qualified trained professional | Trained non-professional | Details not given | | | |
| Endpoint Definition | | | | | | Average Score |
| Endpoint | Accepted and well defined | Equivocal significance | Uncertain significance | | | |
| | | | | | | |
| Results | Clearly reported | Reported | Not reported | | | |
| Dose response relationship | Significant | Marginal | Not examined | | | |
| Confounders/Control | | | | | | Average Score |
| List of potential confounders/covariates | Comprehensive and relevant | Less comprehensive but relevant | Inadequate | | | |
| Major Confounding Factors:<br><br>Age<br>Sex<br>SES<br>Smoking<br>Dietary habits<br>Other occupational carcinogens<br>Ethnicity<br>Arsenic | | | | | | |
| Associations of potential confounders with exposure/effect | Documented for both | Documented for one only | Not documented | | | |
| Assessment of relevant co-exposures | Yes, several | Yes, one | None | | | |

| | |
|---|---|
| **OVERALL TOTAL SCORE** | |
| **Additional Comments:** | |

STUDY ANALYSIS - CARDIOVASCULAR

| | | | | | |
|---|---|---|---|---|---|
| Source<br>Title:<br>Author:<br>Journal:<br>Study Type: prospective – longitudinal – cross-sectional<br>Predominant exposure source (if known): | | | | | |
| NOTE: Studies were considered to be ineligible for analysis if the population size was <50 or the cohort consisted of children <16 years of age. If there were two or more papers testing the same study population, only the publication providing the most detailed information was considered. | | | | | |
| | High = 3 | Medium = 2 | Low = 1 | | |
| Cohort Definition and Size | | | | | Average Score |
| Study population<br>-sample size and power | Sufficient power for anticipated effect | Marginal power for anticipated effect | Inadequate power | | |
| Selection/participation bias<br>☐ Sample recruitment (exclusion/inclusion criteria)<br>☐ Sample characteristics<br>☐ Characteristics (and number/%) of eligible non-participants | Well described | Poorly described | Not described | | |
| Control group:<br>(If applicable) | Present, highly comparable to the exposed group | Present, comparable to the exposed group | Absent | | |
| Exposure | | | | | Average Score |
| Exposure measure | Serial blood-leads<br>Bone Lead | Single concurrent plus additional retrospective marker, as e.g. tooth, dentine | Single concurrent blood Lead | | |
| Exposure history<br>-frequency, duration, intensity | Precisely defined | Reconstructed | Unknown | | |
| Analytical quality control of exposure assessment | Well documented | Roughly documented | Not documented | | |
| Examiner | | | | | Average Score |
| Examiner | Qualified trained professional | Trained non-professional | Details not given | | |
| If several examiners | Inter-observer reliabilities given | Inter-observer reliability monitored but not given | Inter-observer reliabilities not given | | |
| Blindedness | Blinded | Doubtful | Not blinded | | |
| Quality criteria: operator training, standardization of measurement method, errors inherent in equipment used | Adequate | Partial | Inadequate | | |
| Blood Pressure Measurement | | | | | Average Score |
| Systolic and diastolic blood pressure readings (average value) separated by 2 minutes with subject in seated position | >3 readings | 3 readings | Single reading | | |
| Similar blood pressure measurement conditions between exposed and control groups | Yes | Unknown | No | | |
| Measurement technique | Ambulatory | Ausculatory | Self measured | | |
| Application of quality assurance program during data collection | Adequate | Partial | Inadequate | | |

STUDY ANALYSIS – CARDIOVASCULAR (page 2)

| Endpoint Definition | | | | Average Score |
|---|---|---|---|---|
| If assessed, was a clinically accepted definition of hypertension used? | Yes | Unknown | No | |
| For effect measures other than blood pressure and hypertension, is endpoint of established clinical significance? | Established clinical significance with good quality criteria | Less established clinical significance with some quality criteria | Ad hoc measure without quality criteria | |
| Study population: men/women, black/white | Studied as separate groups | Partially studied as separate group | Not studied | |
| Results | Clearly reported | Reported | Not reported | |
| Dose response relationship | Significant | Marginal | Not examined | |

| Confounders/Control | | | | Average Score |
|---|---|---|---|---|
| List of potential confounders/covariates | Comprehensive and relevant | Less comprehensive but relevant | Inadequate | |
| Major Confounding Factors:<br>☐ Age<br>☐ Sex<br>☐ Race<br>☐ Body Mass Index | | Additional Confounders of proven importance:<br>☐ Smoking habit<br>☐ Alcohol consumption<br>☐ SES<br>☐ Menopause Status<br>☐ Diet<br>☐ Blood hematocrit<br>☐ Physical activity<br>☐ Psychological stress<br>☐ Hemoglobin<br>☐ Serum ferritin<br>☐ Drugs | | |
| Associations of potential confounders with exposure/effect | Documented for both | Documented for one only | Not documented | |
| Regression modelling | Described in detail | Poorly described | Not described | |
| Assessment of relevant co-exposures | Yes, several | Yes, one | None | |

| **OVERALL TOTAL SCORE** | |
|---|---|

**Additional Comments:**

STUDY ANALYSIS - NEUROLOGICAL - PEDIATRIC

| Source<br>Title:<br>Author:<br>Journal:<br>Study Type: prospective – longitudinal – cross-sectional<br>Predominant exposure source (if known): | | | | |
|---|---|---|---|---|
| | High = 3 | Medium = 2 | Low = 1 | |
| Cohort Definition and Size | | | | Average Score |
| Study purpose | Precisely defined | Defined without precision | Not defined | |
| Study population<br>-sample size and power | Sufficient power for anticipated effect (>300) | Marginal power for anticipated effect (100-300) | Inadequate power (<100) | |
| Selection/participation bias<br>☐ Sample recruitment (exclusion/inclusion criteria)<br>☐ Sample characteristics<br>☐ Characteristics (and number/%) of eligible non-participants | Well described | Poorly described | Not described | |
| Control group<br>(If applicable) | Present, highly comparable to the exposed group | Present, comparable to the exposed group | Absent | |
| Exposure | | | | Average Score |
| Exposure measure | Serial blood-leads | Single concurrent plus additional retrospective marker, as e.g. tooth, dentine or bone Lead | Single concurrent blood Lead | |
| Analytical quality control of exposure assessment | Well documented | Roughly documented | Not documented | |
| Examiner | | | | Average Score |
| Examiner | Qualified trained professional | Trained non-professional | Details not given | |
| If several examiners | Inter-observer reliabilities given | Inter-observer reliability monitored but not given | Inter-observer reliabilities not monitored | |
| Blindedness | Blinded | Doubtful | Not blinded | |
| Endpoint Definition | | | | Average Score |
| Endpoint | Accepted and well defined | Equivocal significance | Uncertain significance | |
| Effect measure | Established clinical tool with good quality criteria | Less established tool with some quality criteria | Ad hoc measure without quality criteria | |
| Results | Clearly reported | Reported | Not reported | |
| Dose response relationship | Significant | Marginal | Not examined | |
| Confounders/Control | | | | Average Score |
| Maternal IQ considered | Yes | Yes, “proxy” IQ measure | No | |
| Parenting considered | Yes | Partly | No | |
| Maternal mental health considered | Yes | Partly | No | |
| Other measures of home environment considered | Yes | Partly | No | |
| Other potential confounders considered:<br>☐ Age<br>☐ Sex<br>☐ SES<br>☐ Nutrition<br>☐ Ethnicity | Yes | Partly | No | |
| Associations of potential confounders with exposure/effect | Documented for both | Documented for one only | Not documented | |
| Regression modeling | Described in detail | Poorly described | Not described | |
| Assessment of relevant co-exposures | Yes, several | Yes, one | None | |

| **OVERALL TOTAL SCORE** | |
|---|---|
| **Additional Comments:** | |

STUDY ANALYSIS - KIDNEY – GENERAL POPULATION

| | | | | | | |
|---|---|---|---|---|---|---|
| Source<br>Title:<br>Author:<br>Journal:<br>Study Type: prospective – longitudinal – cross-sectional<br>Predominant exposure source (if known): | | | | | | |
| | High = 3 | Medium = 2 | Low = 1 | | | |
| Cohort Definition and Size | | | | | | Average Score |
| Study population<br>-sample size and power | Sufficient power for anticipated effect | Marginal power for anticipated effect | Inadequate power | | | |
| Selection/participation bias<br>☐ Sample recruitment (exclusion/inclusion criteria)<br>☐ Sample characteristics<br>☐ Characteristics (and number/%) of eligible non-participants | Well described | Poorly described | Not described | | | |
| Control group<br>(If applicable) | Present, highly comparable to the exposed group | Present, comparable to the exposed group | Absent | | | |
| Exposure | | | | | | Average Score |
| Exposure measure | Blood-leads | Air, urine, bone or other measure | None | | | |
| Exposure history<br>-frequency, duration, intensity | Precisely defined | Reconstructed | Unknown | | | |
| Analytical quality control of exposure assessment | Well documented | Roughly documented | Not documented | | | |
| Endpoint Definition | | | | | | Average Score |
| Endpoint | Accepted and well defined | Equivocal significance | Uncertain significance | | | |
| | | | | | | |
| Results | Clearly reported | Reported | Not reported | | | |
| Dose response relationship | Examined | | Not examined | | | |
| Confounders/Control | | | | | | Average Score |
| List of potential confounders/covariates | Comprehensive and relevant | Less comprehensive but relevant | Inadequate | | | |
| Major Confounding Factors:<br><br>Age<br>Analgesics<br>Alcohol<br>Diabetes<br>Dietary habits<br>Other occupational nephrotoxins, e.g., Cadmium | | | | | | |
| Associations of potential confounders with exposure/effect | Documented for both | Documented for one only | Not documented | | | |
| Assessment of relevant co-exposures | Yes, several | Yes, one | None | | | |

| | |
|---|---|
| **OVERALL TOTAL SCORE** | |
| **Additional Comments:** | |

## A 1.4. Examples of a checklist used in the Cd-RA to evaluate studies on human health

INSTITUT FÜR
SOZIAL- UND PRÄVENTIVMEDIZIN
DER UNIVERSITÄT ZÜRICH

CH-8006 Zürich,
Sumatrastrasse 30
Telefon 01 / 634 46 11
Telefax 01 / 634 49 86

ABTEILUNG FÜR ARBEITS-
UND UMWELTMEDIZIN

**CROSS-SECTIONAL STUDIES.**

**1. Other publications on the same population.**

Consider possible overlapping with other studies by quite different authors which included part of the same study population.
If relevant : on the first page indications about possible redundancy (part of the same authors have already published results on part of this study population). Indicate references for easy retrieval of publications. Briefly comment on differences / similarities between former and latter pubications and justify the exclusion of the study not considered.

**2. Location.**

Country / region / institution (selection bias !).

**3. Purpose**

precisely defined ?

**4. Study population**

**4.1. and 4.2. Exposed and nonexposed.**

- size of study population, men /women, age. If possible : socioeconomic class, smoking, alcohol, and other relevant factors in that context.
- is the study population young ?
- definition of the study population :
a) beginning of work at...

b) end of work at ...; b) if sequential cross-sectional study or inclusion of retired workers
c) minimal duration;

d) all workers presently working and exposed ? Retired and currently nonexposed workers ?
Time since last exposure ?

Are workers previously diagnosed with poisoning included ?

**4.3. Final study population.**

- initial vs final study population (as a summary showing the lost cases and controls at each step of constitution of study population with the most important data on age, sex, employablity, and other important variables in that context).
- comparability of cases and controls as for : age, sex, socioeconomic group, education, and other important variables in that context.
Are workers previously diagnosed with poisoning included ?
IMPORTANT : considerable differences may be found between initial and final study population. A presentation of the results (for example in tables) should take this issue into account.

**5. Selection, participation rate, representativeness.**

- selection of the study population
- selection of the study sample
- participation rate
- representative sample ?

If sequential cross-sectional study : are the samples comparable ?

**6. Exposure.**

**6.1. Specific aspects : exposed workers**

Is the word "exposed" clearly defined (with respect to type, minimal intensity, duration, frequency ?). Observational period (if not mentioned under 4.; important because of time-related changes of exposure intensity) ? Previous poisonings ?

- type :
== general population or industry ?
== type of industry (for example Cd exposure : cadmium production, alloys, soldering and/or cutting, Cd-Ni battery, etc.) ?
== is "exposure" defined by occupation and/or industry, group of agents, agent ? Are the groups specific or very broad (= how specific if this definition ?) ? Are concomitant exposures possible (for example : heavy metals vs Cd + As inorg or Pb or Ni vs Cd only ? Benzene in garages, oil refineries, printing plants represent three quite different exposure conditions).

- information on exposure frequency, duration, and intensity :
== yes/no
== only present exposure (strictly cross-sectional) or information on previous exposure (in this plant, in the same occupation but in other plants, in all occupations for the lifetime)
== minimal intensity, duration, frequency : based on exposure reconstruction or objective measures ?
== if exposure reconstruction : type of variable (dichotomous if exposed vs nonexposed, ordinal, exposure score, etc.) ?
== if dichotomous classifications : is the cut-off clearly described, credible, arbitrary ? Are minimal intensity, duration, frequency taken into account to define the word "exposed" ?
== if ordinal categories or of exposure score : is the classification / score credible, consistent ? Is there any indication of the validity of the classification / score ?
== objective measures available ? air sampling (area vs. personal, total vs respirable dust); biol.monitoring (blood/urine/neutron activation analysis/x-ray fluorescence, etc.)

samples from controls and exposed workers examined in the same series ?
quality control (exposure assessment) ?

**6.2. Specific aspects : control workers**

**6.3. Summary**

**7. Diagnosis.**

Is the endpoint clinically relevant (predictive value) ? Methods ? Quality control ?

If relevant :
Classification scheme
Are there objective criteria required for ascertaining diagnosis (example : FAB, SLE) ?
Blind review of medical records, slides, if any ?
Panel review ?
Other important methodologic aspects (example : biopsy for kidney diseases, immunofluorescence for glomerulonephritis, histolological confirmation for cancer).

**8. Bias.**

- preplacement examination
- healthy worker effect
- is it clear that the endpoint is really an effect of the exposure (cross-sectional design !)

**9. Interview and coding, laboratory.**

Blind interview / interview procedure / structure and content of the interview.
Blind coding of the answers / coding according to .... (are criteria mentioned, credible, arbitrary).

INSTITUT FÜR
SOZIAL- UND PRÄVENTIVMEDIZIN
DER UNIVERSITÄT ZÜRICH

CH-8006 Zürich,
Sumatrastrasse 30
Telefon 01 / 634 46 11
Telefax 01 / 634 49 86

ABTEILUNG FÜR ARBEITS-
UND UMWELTMEDIZIN

## RETROSPECTIVE COHORT STUDIES. MORTALITY.

### 1. Other publications on the same population.

. Consider possible overlapping with other studies by quite different authors which included part of the same study population. If relevant : on the first page indications about possible redundancy (part of the same authors have already published results on part of this study population). Indicate references for easy retrieval of publications. Briefly comment on differences / similarities between former and latter pubications and justify the exclusion of the study not considered.

### 2. Location.

Country / region / institution.

### 3. Purpose

precisely defined / hypotheses - generating

### 4. Study population

### 4.1. and 4.2. Exposed and nonexposed.

- cohort size, men /women, age, number / percentage of deaths (obs. /exp. numbers of deaths). If possible : socioeconomic class, smoking, alcohol, and other relevant factors in that context.
- is the percentage of deaths higher than 10 % ?
- is the cohort young ?
- definition of the cohort :

a) begin of work at...
b) end of work at ...
c) minimal duration
d) other characteristics

Definition of follow - up : begin / end of follow - up
Are workers previously diagnosed with poisoning included ?

### 4.3. Final study population.

- initial vs final study population (as a summary showing the lost cases and controls at each step of constitution of study population with the most important data on age, sex, employablity, and other important variables in that context).
- comparability of cases and controls as for : age, sex, socioeconomic group, education, and other important variables in that context.

Are workers previously diagnosed with poisoning included ?
IMPORTANT : considerable differences may be found between initial and final study population. A presentation of the results (for example in tables) should take this issue into account.

### 5. Selection, participation rate, representativeness.

- selection
- participation rate
- representative sample

If register : is the coverage good ?
If morbidity / mortality statistics : data quality ?

### 6. Exposure.

### 6.1. Specific aspects.

Is the word "exposed" clearly defined (with respect to type, minimal intensity, duration, frequency ?). Observational period (if not mentioned under 2.; important because of time-related changes of exposure intensity) ? Previous poisonings ?

- type :
== general population or industry ?
== type of industry (for example : exposure to heavy metals/to Cd + As inorg or Pb or Ni/or to Cd only ? Benzene in garages, oil refineries, printing plants represents three quite different exposure situations).

== is "exposure" defined by occupation and/or industry, group of agents, agent ? Are the groups specific or very broad (= how specific if this definition ?) ? Are concomitant exposures possible (for example : heavy metals vs Cd + As inorg or Pb or Ni vs Cd only ? Benzene in garages, oil refineries, printing plants represent three quite different exposure conditions).
== if coding of occupations : clearly standardized ? Based on which coding system (for example : Dictionary of Occupational Titles of the Census) ? Blind ?

- information on exposure frequency, duration, and intensity :
== yes/no
== minimal intensity, duration, frequency : based on exposure reconstruction or objective measures ?
== if exposure reconstruction : type of variable (dichotomous if exposed vs nonexposed, ordinal, exposure score, etc.) ?
== if dichotomous and based on death certificates, registers, or similar sources of information : longest, usual, current, last occupation or occupation at diagnosis ?
== if other dichotomous classifications : is the cut-off clearly described, credible, arbitrary ? Are minimal intensity, duration, frequency taken into account to define the word "exposed" ?
== if ordinal categories or of exposure score : is the classification / score credible, consistent ? Is there any indication of the validity of the classification / score ?
== objective measures available ? air sampling (area vs. personal, total vs respirable dust); biol.monitoring (blood/urine/neutron activation analysis/x-ray fluorescence, etc.)

**7. Diagnosis.**

Classification scheme (ICD, ICD - O, etc.)
Are there objective criteria required for ascertaining diagnosis (example : FAB, SLE) ?
Blind review of medical records, slides, if any ?
Panel review ?
Other important methodologic aspects (example : biopsy for kidney diseases, immunofluorescence for glomerulonephritis, histolological confirmation for cancer).
If death certificates : underlying vs. contributing cause of death.
- high / low mortality rate ?

**8. Bias.**

- surveillance bias
- changes in the course of the study (for example : job changes in comparison to the job used as exposure surrogate)
- diagnostic access bias
- diagnostic suspicion bias

**9. Interview and coding.**

Blind interview / interview procedure / structure and content of the interview.
Blind coding of the answers / coding according to .... (are criteria mentioned, credible, arbitrary).

**10. Design and statistics.**

- design
- SIR, SMR, PMR
- reference population : national, regional, other industrial workers, low vs high exposure
- statistical methods

**11. Confounding factors**

Age, sex, hospital, smoking, alcohol ?
If relevant : socioeconomic group, residence, genetic / familial factors, race, ethnicity

Considerable sources of misclassification ? (for exposure and disease see 6. and 7., respectively)

Sensitivity analysis ?

IMPORTANT : were the confounding factors taken into account in the analysis or were they only mentioned as items in the interview and not considered in the statistical analysis ? Are these factors clearly defined (nationality may change after wedding) ? If subgroups are used are these subgroups relevant ?

**12. Results.**
**12.1. Results**
**12.2. What about power ?**

**13. Identification, latency, DRC.**

Identification : of a specific causal agent, specific causal occupation ?
Latency time (lagging of some years) : yes / no ? biologically credible ?
Dose - response curve : was it examined ?

**14. Physiopathology.**

Physiopathologial mechanisms (plausibility)

**15. Miscellaneous.**

26.9.1997

**A 1.5. Toxicological study reliability code documentation (Source: NiPERA)**

Interpretation guidance describing the way the overall reliability code is derived is given on the next two pages.

CAS #: _____ Reviewers Initials: _____ RELIABILITY CODE:_see footnote_

Date: _____ RELEVANT: Y N

Study ID: _____ Endpoint Evaluated: _____

(Report # or Archive Code)

| Acceptability Criteria | Yes | No | NA* | NR* | Comments |
|---|---|---|---|---|---|
| Guideline study (OECD/EPA/EC/ MITI, etc.) Or equivalent study meets national/scientific stds. | | | | | |
| GLP or equivalent followed | | | | | |
| Documentation sufficient for assessment | | | | | |
| Appropriate group sizes and sex | | | | | |
| Appropriate dose levels | | | | | |
| Appropriate concurrent control group(s) /responses | | | | | |
| Appropriate historical controls, if needed | | | | | |
| Appropriate statistics | | | | | |
| Appropriate analytical method (test compound) | | | | | |
| Analytical verification of test material & dose concentrations | | | | | |
| Data from handbook or collection of data | | | | | |
| Accepted calculation method | | | | | |
| Abstract or secondary citation | | | | | |
| | | | | | |
| Other relevant info on back of form | | | | | |
| **Significant Study Defects - list below** | **Yes** | **No** | **NA*** | **NR*** | **Comments** |

COMMENTS (ie relevance - dose, species, route of exposure etc.):

**Possible reliability codes:**

**1 = Valid without restriction**
1a: GLP guideline study
1b: Comparable to guideline study
1c: Meets national standard methods (AFNOR/DIN)
1d: Meets generally accepted scientific standards and is described in sufficient detail

**2 = Valid with restriction**
2a: Guideline study without detailed documentation
2b: Guideline study with acceptable restrictions
2c: Comparable to guideline study with acceptable restrictions
2d: Meets national standard methods with acceptable restrictions
2e: Meets generally accepted scientific standards, well documented and acceptable for assessment

2g: Data from Handbook or collection of data

**3 = Invalid**
3a: Documentation insufficient for assessment
3b: Significant methodological deficiencies
3c: Unsuitable test system

**4 = Not Assignable**
4a: Abstract
4b: Secondary literature
4c: Original reference not yet available
4d: Original reference in foreign language
4e: Documentation insufficient for assessment

* NA = Not applicable for this study type NR = Not reported

**TOXICOLOGY STUDY RELIABILITY CODE DOCUMENTATION- Interpretation Guidance**

CAS #: _____ Reviewers Initials: _____ RELIABILITY CODE: 1 and 2

Date: _____ Relevant Y N

Study ID:_____ Endpoint Evaluated: _____

(Report # or Archive Code)

| Acceptability Criteria | Yes | No | NA* | NR* | Comments |
|---|---|---|---|---|---|
| Guideline study (OECD/EPA/EC/ MITI, etc.) | x | | | | Required for 1a; 2a,b |
| Or equivalent study meets national/scientific stds. | x | | | | Required for 1 b,c,d; 2c,d,e |
| GLP or equivalent followed | x | | | | Required for 1a |
| Documentation sufficient for assessment | x | | | | Required for 1a,b,c,d and 2a,b,c,d,e |
| Appropriate group sizes and sex | | | | | Apply judgement for impact on rating |
| Appropriate dose levels | | | | | A |
| Appropriate concurrent control group(s) /responses | | | | | A |
| Appropriate historical controls, if needed | | | | | A |
| Appropriate statistics | | | | | A |
| Appropriate analytical method (test compound) | | | | | A |
| Analytical verification of test material | | | | | A |
| & dose concentrations | | | | | A |
| Data from handbook or collection of data | x | | | | Supplemental data only - 4 (2-acute) |
| Accepted calculation method | x | | | | Supplemental data only - 4 |
| Abstract or secondary citation | x | | | | Supplemental data only - 4a |
| | | | | | |
| Other relevant info on back of form | | | | | |
| **Significant Study Defects - list below** | **Yes** | **No** | **NA*** | **NR*** | **Comments** |
| | x | | | | Cannot be checked for 1 or 2 |

COMMENTS (ie relevance - route of exp., species, dose)

**Possible reliability codes:**

**1 = Valid without restriction**
1a: GLP guideline study
1b: Comparable to guideline study
1c: Meets national standard methods (AFNOR/DIN)
1d: Meets generally accepted scientific standards and is described in sufficient detail

**2 = Valid with restriction**
2a: Guideline study without detailed documentation
2b: Guideline study with acceptable restrictions
2c: Comparable to guideline study with acceptable restrictions
2d: Meets national standard methods with acceptable restrictions
2e: Meets generally accepted scientific standards, well documented and acceptable for assessment

2g: Data from Handbook or collection of data

**3 = Invalid**
3a: Documentation insufficient for assessment
3b: Significant methodological deficiencies
3c: Unsuitable test system

**4 = Not Assignable**
4a: Abstract
4b: Secondary literature
4c: Original reference not yet available
4d: Original reference in foreign language
4e: Documentation insufficient for assessment

* NA = Not applicable for this study type NR = Not reported

**TOXICOLOGY STUDY RELIABILITY CODE DOCUMENTATION - Interpretation Guidance**

CAS #: _____ Reviewers Initials: _____ RELIABILITY CODE: 3 and 4

Date: _____ Relevant: Y N

Study ID: _____ Endpoint Evaluated: _____

(Report # or Archive Code)

| Acceptability Criteria | Yes | No | NA* | NR* | Comments |
|---|---|---|---|---|---|
| Guideline study (OECD/EPA/EC/ MITI, etc.) | | x | | | If both are No requires 3 |
| Or<br>equivalent study meets national/scientific stds. | | x | | | |
| GLP or equivalent followed | | | | | |
| Documentation sufficient for assessment | | x | | | Requires 3 or 4 |
| Appropriate group sizes and sex | | | | | A |
| Appropriate dose levels | | | | | A |
| Appropriate concurrent control group(s) /responses | | | | | A |
| Appropriate historical controls, if needed | | | | | A |
| Appropriate statistics | | | | | A |
| Appropriate analytical method (test compound) | | | | | A |
| Analytical verification of test material | | | | | A |
| &<br>dose concentrations | | | | | A |
| Data from handbook or collection of data | x | | | | Supplemental data only - 4 |
| Accepted calculation method | x | | | | Supplemental data only - 4 |
| Abstract or secondary citation | x | | | | Supplemental data only - 4a,b |
| | | | | | |
| Other relevant info on back of form | | | | | |
| **Significant Study Defects - list below** | **Yes** | **No** | **NA*** | **NR*** | **Comments** |
| | x | | | | Requires 3 |

**Possible reliability codes:**

**1 = Valid without restriction**
1a: GLP guideline study
1b: Comparable to guideline study
1c: Meets national standard methods (AFNOR/DIN)
1d: Meets generally accepted scientific standards and is described in sufficient detail

**2 = Valid with restriction**
2a: Guideline study without detailed documentation
2b: Guideline study with acceptable restrictions
2c: Comparable to guideline study with acceptable restrictions
2d: Meets national standard methods with acceptable restrictions
2e: Meets generally accepted scientific standards, well documented and acceptable for assessment

2g: Data from Handbook or collection of data

**3 = Invalid**
3a: Documentation insufficient for assessment
3b: Significant methodological deficiencies
3c: Unsuitable test system

**4 = Not Assignable**
4a: Abstract
4b: Secondary literature
4c: Original reference not yet available
4d: Original reference in foreign language
4e: Documentation insufficient for assessment

* NA = Not applicable for this study type NR = Not reported