

**There is a saying that
when 5 experts meet
there are 7 opinions,....**

**...but this is not true for toxicologists,
is it?**

Toxicological Data Reliability Assessment Tool

ToxRTool

ECVAM project: Development of a quality assessment tool for toxicological data

Klaus Schneider

FoBiG, Forschungs- und Beratungsinstitut Gefahrstoffe GmbH
Freiburg, Germany

Quality of toxicological

Scientifically: which data are acceptable to cover an endpoint?
Monetarily: which study sells?

- Chemicals legislation (formerly: existing chemicals, HPV programs, now: REACH)
- Data used for Classification and Labelling (GHS)
- Validation of alternative methods: retrospective validation using existing data, validation against existing data

SOT 2007:
lacking system
for quality
evaluation is
obstacle
to harmonised
classification

Klimisch et al. 1997

Problem in IUCLID5:
only reliability
assessment is foreseen

Defined

Reliability: inherent quality of data

Relevance: appropriateness for a particular hazard identification or risk characterisation

Adequacy: usefulness for risk assessment purposes

Klimisch categories: widely used for categorising reliability

- 1: reliable without restriction
- 2: reliable with restrictions
- 3: not reliable
- 4: not assignable

Scope and objectives of the ECVAM project

- Develop a tool for assessing the reliability of toxicological data (in vitro, in vivo)

- The tool is expected to
 - improving transparency of reliability assessments and
 - to provide guidance for a harmonised approach (leading to more homogenous assessments)

- It is not an objective to achieve homogenous evaluations when there are differences in judgements

What is the “tool”

- Consists of two parts: in vitro and in vivo list of criteria
- List of criteria (with explanations) to be answered by yes or no
- Red criteria: minimum information requirements (substance identified, species identified, ...)
- Implemented as Excel worksheets
- Added scores lead to Klimisch 1, 2 or 3
(tool not applicable to category 4: secondary sources)
- Note: All reports/publications are treated equally – no bonus for guideline studies

Guideline studies are expected to be Cat 1, but a short report on guideline study (without detailed documentation) might be Cat 3

ToxRTool: in vivo

Reliability assessment of in vivo toxicity studies

Remarks for study under evaluation:
###

Authors:

Titel:

Testing facility, year, sponsor, study no. or bibliographic reference:

Explanations are available for most criteria and show up, when the cursor is moved over the criteria field. Please read carefully!
Red criteria: a score of 1 is needed for these criteria to achieve reliability category 1 or 2 (see worksheet Explanations): Please evaluate with special care!

Criteria			Evaluator's explanation
No.	Criteria Group I: Test substance identification	Score	
1	Was the test substance identified?		
2	Is the purity of the substance given?		
3	Is information on the source/origin of the substance given?		
4	Is all information on the nature and/or physico-chemical properties of the test item given, which you deem indispensable for judging the data (see explanation for examples)?		
		0	

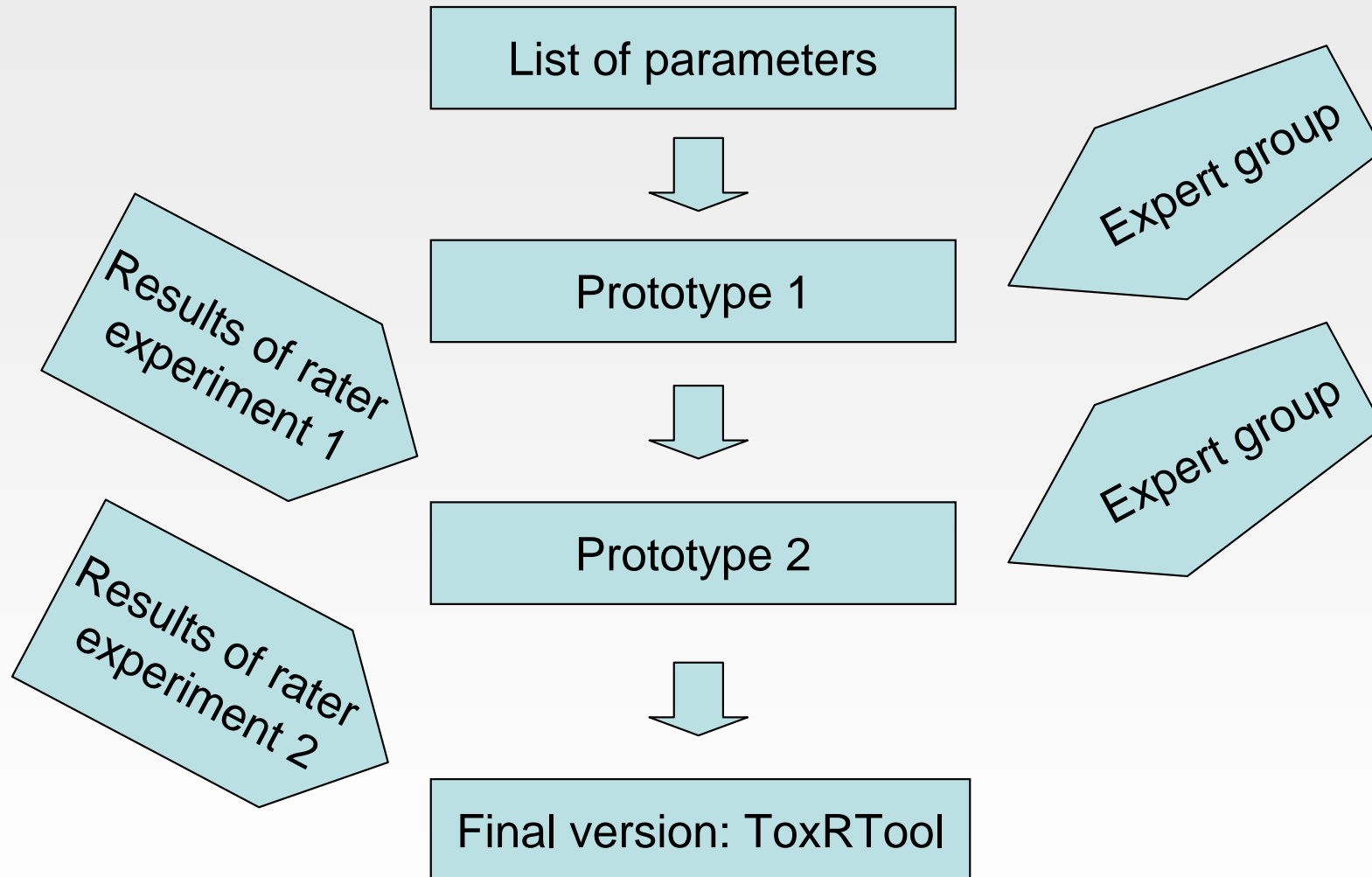
	A	B	C	
40		Criteria Group IV: Study results documentation		
41	17	Are the study endpoint(s) and their method(s) of determination clearly described?		
42	18	Is the description of the study results for all endpoints investigated transparent and complete?		
43	19	Are the statistical methods applied for data analysis given and applied in a transparent manner (give also point, if not necessary/applicable, see explanations)?		
44			0	
45				
46		Criteria Group V: Plausibility of study design and results		
47	20	Is the study design chosen appropriate for the study objective (see explanations for details)?		
48	21	Are the quantitative study results reliable (see explanations for arguments)?		
49			0	
50				
51			0	
52				
53		A Numerical result leads to initial Category:	3	
54		B Checking red scores leads to revised Category:	3	
55		C Evaluator's proposal: Category:		
56		D Justification in case evaluator deviates from B:		
57				
58				
59		Optional documentation of observations with importance to relevance		
60		During the course of the quality assessment observations may be made which are important for discussing the relevance of the data for specific purposes. The optional possibility is provided here to document these observations for future use.		
61		What is the purpose of the quality evaluation performed (data documentation for use under REACH, classification activity under GHS, ECVAM validation activities, other)?		
62		Testing the tool (part of rater experiment 2)		

Project frame

- Project duration from April 2007 – October 2008
- Prototype 1 developed from list of influencing parameters
- Tested in two rater experiments
- Accompanying expert group

Name	Affiliation
Neil Carmichael	ECETOC AISBL Brussels, Belgium
Karl-Heinz Cohr, on behalf of EUROTOX, ERAS	DHI Water • Environment • Health Hørsholm, Denmark
Cees de Heer	RIVM Centre for Substances and Integrated Risk Assessment Bilthoven, The Netherlands
Sebastian Hoffmann (coordinator, chair)	European Commission, JRC, Institute of Health and Consumer Protection (IHCP), ECVAM, Ispra (VA), Italy
Dinant Kroese	TNO Quality of Life Zeist, The Netherlands
Franz Oesch, on behalf of EUROTOX, ERAS	Institute of Toxicology, University of Mainz Mainz, Germany
Iona Pratt	Food Safety Authority of Ireland, Dublin, Ireland
Hans-Bernhard Richter-Reichhelm	Bundesinstitut für Risikobewertung (BfR) Berlin, Germany

Approach



Set-up of rater experiments

Rater experiment 1:

- Prototype 1: 25 criteria for in vitro and in vivo part each
- 11 case studies for in vitro and in vivo part each
- 11 raters for in vitro part and 9 raters for in vivo part

Rater experiment 2:

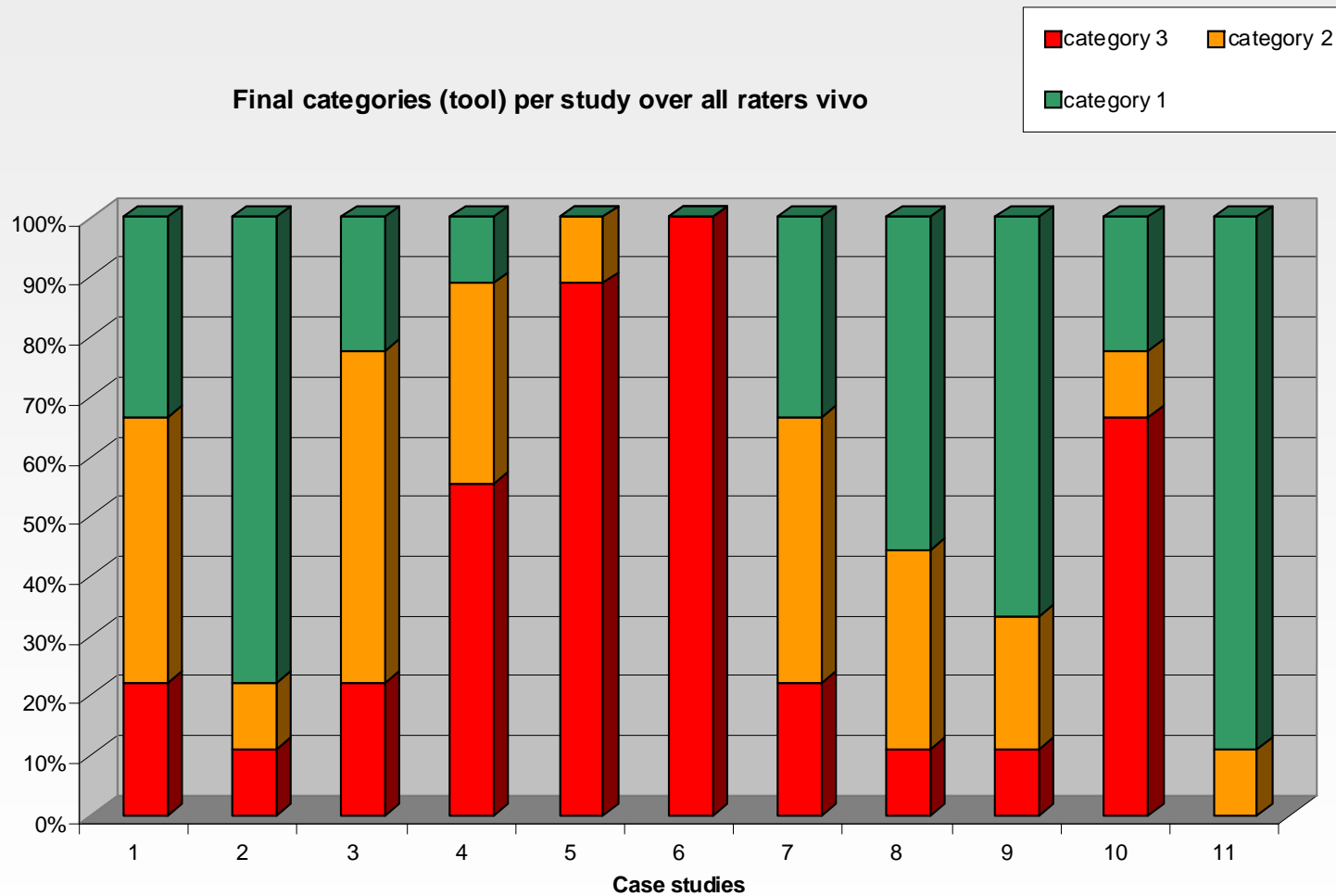
- Prototype 2: 18 criteria for in vitro and 21 for in vivo part
- Same 11 case studies for in vitro and in vivo part each
- 17 raters for in vitro part and 12 raters for in vivo part

Case studies

- 11 case studies for in vitro and in vivo each (same studies for both rater experiments)
- Broad spectrum of endpoints and study types:
 - in vivo: acute tox., genotox., reproduct. Tox., carcinogenicity (standard and knock-out), short term inhalation, experimental human study, toxicokinetic study, skin irritation and sensitisation
 - in vitro: genotoxicity, cytotoxicity, skin penetration, skin corrosion, hepatotoxicity (liver slices), phototoxicity
- Broad spectrum of quality (guideline study, detailed peer-reviewed publications, old data from 50ies, short reports, studies on large lists of substances)

Results of rater experiment 1

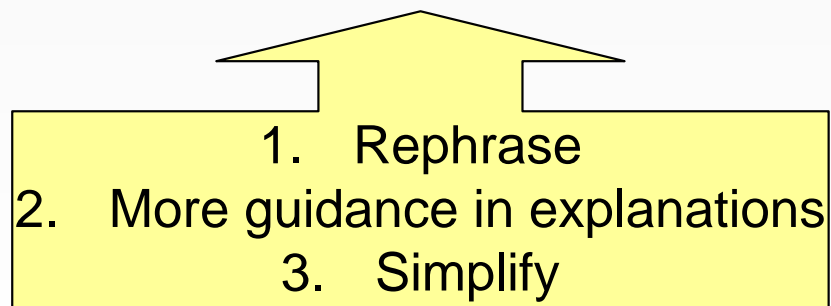
- High variability between raters observed (less with in vivo studies)



Results of rater experiment 1

- Sources of variability:
 - I individual errors or explanations not considered
 - II differentiation between reliability and relevance not followed
 - III diverging interpretation or weighing of available information
 - IV criteria misleading or not applicable to all study types

- Certain criteria can be identified as major sources of variability



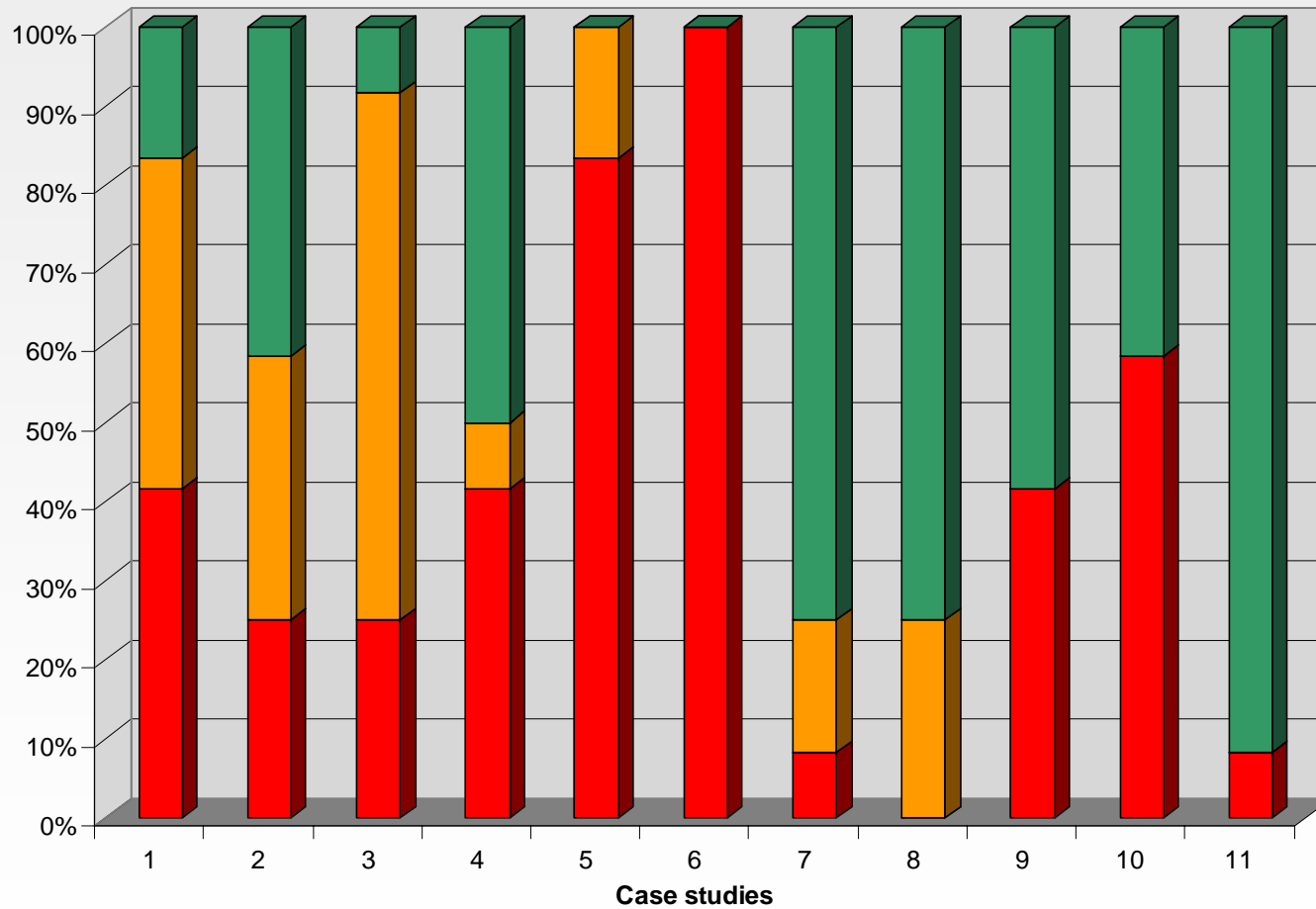
Results of rater experiment 2: General

- Vast majority of raters find the tool useful, user-friendly and transparent
- Raters feel well guided by the tool, even where high variability between raters is observed
- Results of experiment 2 are (somewhat) more homogenous than those of experiment 1, but substantial variability prevails
- Numerical scores are more homogenous than categories
- Red criteria (minimum information set) are important source of variability in categories
- Points of discussion by raters:
 - o Guideline status
 - o Weighing of plausibility criteria

Results of rater experiment 2 - in vivo: Results B

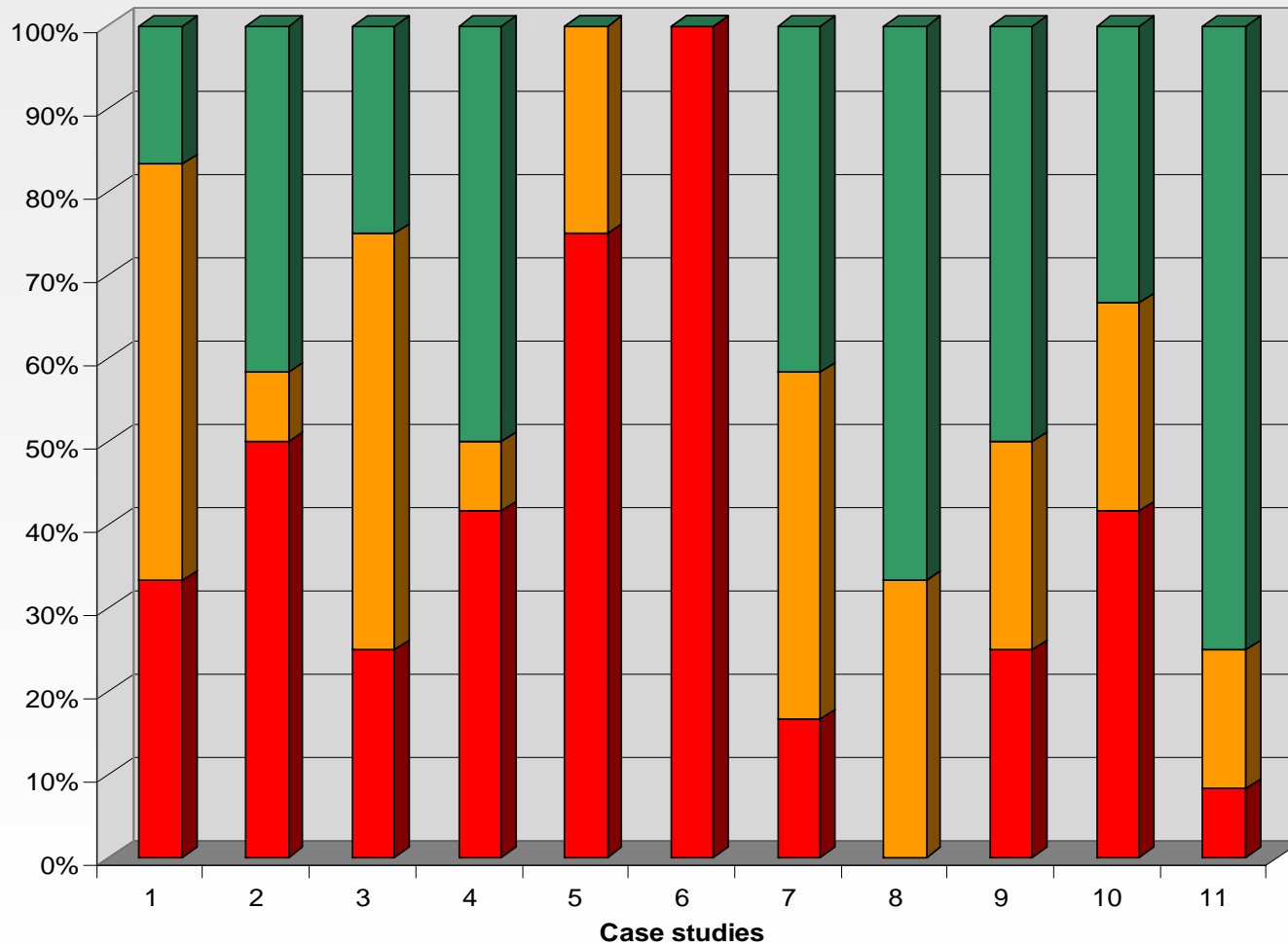
Final categories (tool) per study over all raters vivo

■ category 3
 ■ category 2
 ■ category 1

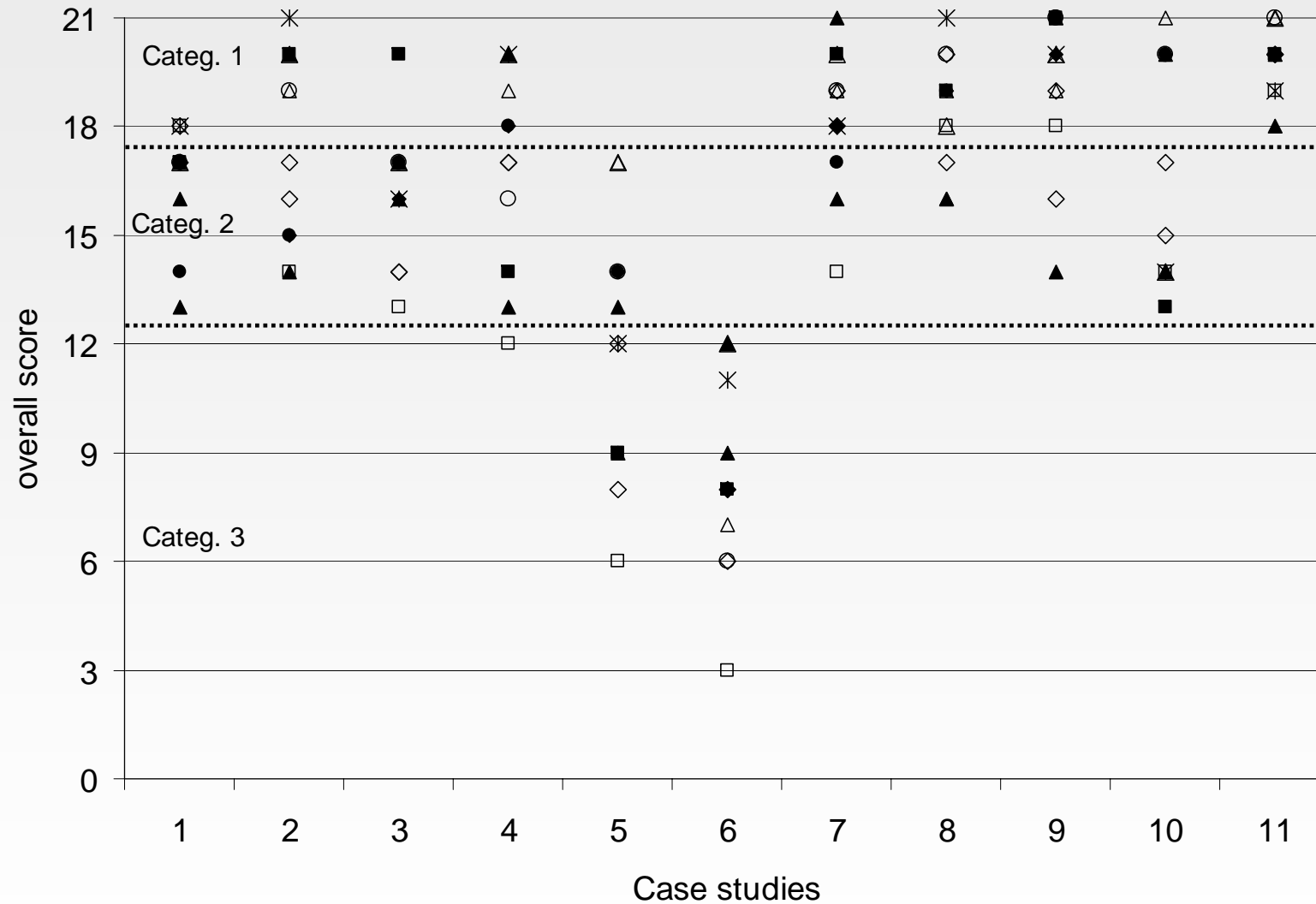


Rater experiment 2 - Results C (raters' own judgement)

Rater's proposal per study over all raters vivo



Results of rater experiment 2 - in vivo: total score



Conclusions

- Quality assessments performed by individual toxicologists tend to vary considerably
- Higher variability is observed with in vitro studies compared to in vivo studies, presumably due to missing guidelines for a lot of in vitro study types
- The assessment tool gives guidance on the “what” and “how” of the assessment
- The assessment tool is able to reduce some, but not all of the (scientifically unjustified) variability
- Practical implementation of the minimum data set (red criteria) is difficult
- The assessment tool is helpful to understand sources of heterogeneity

- Overall conclusion:
 - Tool application is a very helpful exercise
 - Final tool version (in vivo) should go into testing for day-to-day application

Many thanks to

- ❖ ECVAM for funding the project and for Sebastian Hoffmann for his support as coordinator and chair of the expert group
- ❖ The members of the expert group:
Neil Carmichael, Karl-Heinz Cohr, Cees de Heer, Dinant Kroese, Franz Oesch, Iona Pratt, Hans-Bernhard Richter-Reichhelm
- ❖ All raters
- ❖ The project partners from the DKFZ: Annete Kopp-Schneider, Iris Burkholder, Lutz Edler)
- ❖ and my FoBiG colleagues Markus Schwarz and Martin Hassauer

Back-up slides

“Other” sources of variability

1 Individual errors

- Examples: particle size requested for 1,3-butadiene inhalation study
- Substance not identified (e.g. styrene in list of substances)

2 Borderline cases

- Examples: controls only shown in figure, not mentioned
- Controls mentioned but not specified

3 Different views

- Statistical analysis required/not required for a specific study, study design adequate or not (how many individuals, how many test concentrations required for a human experimental study?), etc.
- Guideline status

Red criteria not treated more carefully than others!

Guideline status of studies

Example: case study of Farr et al. 2001

(short report on reproductive toxicity guideline study)

Comment by one rater, who graded study up to Klimisch category 1:

„There is a fully guideline conform and GLP study behind these data which are summarized at a very short level. For strictly regulatory purposes the paper can serve as supplemental evidence only, however if indeed the basic study is fully guideline and GLP compliant, the underlying report (from Hazleton) would be fully usable for a registration”

Evaluation by the tool:

→ Short publication is Category 2 or 3, test report would be 1