Table of Contents

Scientific findings as a basis for decision-making:

Guideline for assessing the due diligence of scientific rationale

(As of 29/04/2019)

Motivation:

Opinions and decisions are often based on scientific findings. However, it is noticeable that in many, if not most cases, seemingly contradictory findings find themselves opposing each other.

This guideline should make it possible to prevent any consideration of apparently untrustworthy bogus findings on the basis of fundamental criteria that are always valid.

The authors are aware that simple criteria do not make it possible to recognise incorrect data that is carefully presented. Professionally manipulated data deliberately hides a more subtle intent, while information that is compiled in a negligent and shoddy manner is often easier to identify.

Indications of manipulation or biased handling of data are usually difficult to recognise within the narrow context of a specific topic. More insight can only provide a fleeting view of the overall context[1,2]. Ultimately, this will often turn out to be a test for criminal intent carried out by higher authorities. Those who fail to perform this test quickly run the risk of trusting the seemingly obvious instead of facts.

The intention of this guideline is to help make a rapid distinction between relevant and non-relevant information that is declared to be "scientific". This approach should also provide a competent, targeted process to uncover data manipulation and the resulting interpretations. It should also make it possible to handle supposedly scientifically sound statements rationally and critically as well as to answer the simple question: "Is that possible?"

Preliminary consideration:

Scientific findings follow a cycle that usually leads from observations and the resulting conjectures about correlated quantities to the formation of hypotheses about causal interdependencies. The essence of the scientific work is then to translate these hypotheses from correlations into scientific theories about causal quantitative relationships. Every scientific theory must be able to withstand independent reviews within its scope at any time and enable measurable quantitative prognoses – a feature referred to as reproducibility (see below). The basis and prerequisite for carrying out independent reviews is the detailed publication of the theories and results of experiments or evaluations.

[1] http://www.faz.net/aktuell/wirtschaft/diesel-affaere/deutsche-umwelthilfe-bekommt-geld-von-toyota-14256098.html

[2] http://www.faz.net/aktuell/wirtschaft/diesel-affaere/deutsche-umwelthilfe-die-diesel-hasser-14246048.html

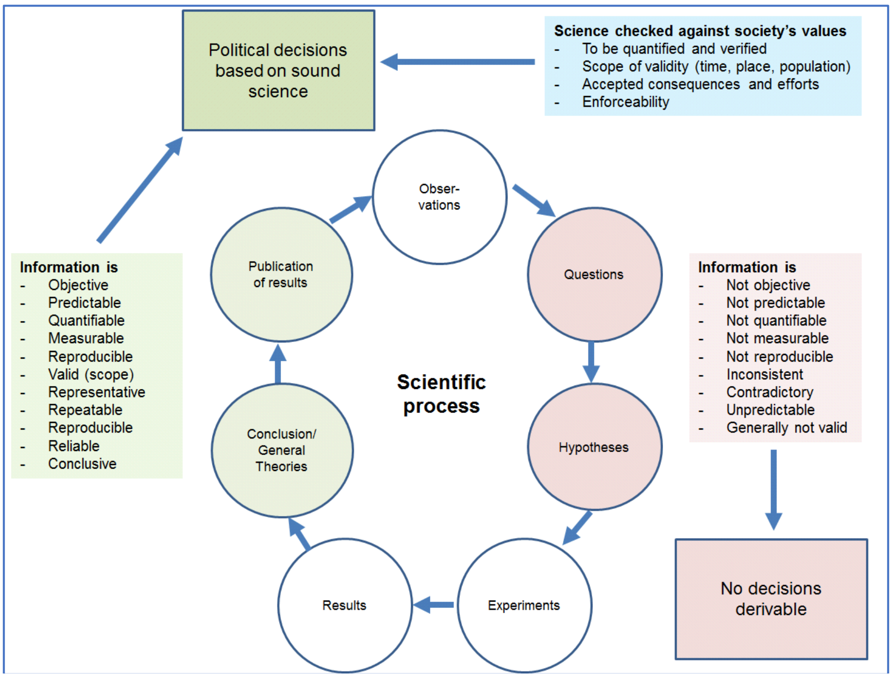

Figure 1: Process to advance scientific knowledge

This paper assumes that a political decision should be based only on interdependencies that enable reliable, quantitative and always valid i.e. clear statements and prognoses regarding the consequences of the decision. Accordingly, pure correlations without knowledge of the quantitative interdependence based on a rationally and objectively determined mechanism are not suitable as a basis for political decisions.

Overview:

Meticulous scientific work creates clear statements about the following criteria of a concrete, well-defined problem:

1. Reproducibility: Are the statements of the scientific theory generally applicable, or are they restricted in terms of space, time, people, or otherwise?
2. Informative value: Does the presentation provide reliable statements about the causalities of the given issue? Is quantifiability possible as a basis for checking predictions when testing?
3. Representativeness: Can the presented information be used to assess the entire scope of the issue, or is the scope more limited?
4. Correctness: Do the results of the scientific theory correctly and quantitatively reflect the causalities, i.e. are systematic errors excluded?
5. Precision: How large is the scattering of the quantitative data of the scientific theory, i.e. how high are random deviations and is it possible to distinguish the determined values from any background noise with sufficient significance?

Hierarchical assessment criteria

Reproducibility:

Are the statements of the scientific theory generally applicable, or are they restricted in terms of space, time, people, or otherwise?

Investigations, measurements and evaluations are inevitably subject to specific influences that can significantly change and falsify the result. Such influences can arise from local-spatial, temporal, personal and many other considerations. Only by independently (in a political, social, personal, financial sense etc.) repeating the scientific study under altered conditions (different place, time, processors, etc.) can it be shown whether the results are generally valid or whether they only occur under special conditions. In the latter case, no reasoned decision of general significance can possibly be made.

Special consideration should also be made of so-called meta-studies. Here, the results from various sources are compared, collated or juxtapositioned. Each study must be independently confirmed.

Informative value:

The prerequisite for making reliable conclusions to allow decisions to be made is a specific issue. What is the subject of the study? What or who should be considered? What should not be considered? If an issue requiring decisions and conclusions is not clarified clearly along with its limits, it will not be possible to make any statements. Its presentation will then be unsuitable for providing reliable statements about causalities, since the subject of the study would be undefined. For this reason, the priority is to define the scope of the study as well as any derived conclusions clearly.

The derived conclusions must be quantifiable. Otherwise it will not be possible to check the results or predictions.

Extrapolations are outside the scope of validity per se and should only be considered as hypothetical, demanding a separate handling.

Again, special attention should be paid to meta-studies or reviews. If the compiled information originates from the assessment of differing hypotheses/issues, this will generally result in misinterpretations. Due to the different focuses or examined aspects, individual studies often fail to meet the essential criteria for assessment within a meta-study.

Representativeness:

After defining the scope of validity of the planned statements and the related studies, it can and must be judged whether the underlying information lies within this scope. Otherwise the scope of validity will need to be restricted or indeed extended. If the originally defined scope of validity is to be retained, additional information may need to be provided. If this is not possible, the study currently being assessed can not be considered as applicable for the topic in question. Even for studies that go beyond the topic in question and assess overarching interdependencies, a seemingly obvious application to a subset of these interdependencies should be carefully considered.

Again, meta-studies or compilations of results should be considered particularly critically. The scope of validity must be appropriate and be clearly identified. Otherwise, comparisons or collations must be considered as inadmissible and worthless.

It is quite difficult to give a positive example for this criterion. Falsification (see negative example) is simple, since only one conflicting example suffices. Verification of this is often uncertain since the parameters are rarely known in their entirety. We therefore provide an overview of a number of conditions that are absolutely required for representativeness.

Correctness:

Systematic errors include following false assumptions, unaccounted for or unknown parameters, as well as unsuitable investigation or evaluation methods. Extrapolations in particular can easily cause systematic errors.

Systematic errors appear as deviations in a certain direction – they always cause values that are either too low or too high.

Systematic errors can only be eliminated by ascertaining the cause of the error and eliminating it; however, this requires a change in the investigation method, as they will merely occur again if the study is simply repeated.

It is difficult to detect systematic errors because the causes are hidden. They are often implicit, unformulated assumptions or parameters. Their validity may be tacit or unconscious – or forgotten or overlooked.

Publishing studies that will be controlled by other scientific working groups is one of the most promising ways to identify and eliminate systematic errors. In doing so, it is decisive that the other working groups achieve completely identical results independently of each other. Often one sees that apparent further developments of argumentations are based on the permanent adoption of the original studies, statements, and beliefs.

A good way to find systematic errors is to perform a plausibility check. This check confronts the apparent results in different contexts or with other measured values than those observed in the study. If the scope of validity is not excluded, then implausible or unrealistic results or statements may indicate systematic deviations.

Precision:

Each study or measurement has random variations. Their causes lie in uncontrollable, yet effective mechanisms. Characteristically, both positive and negative deviations occur. Random deviations can be quantified by mathematical-statistical methods and are an indispensable component of the specification of a value in the form

expected value = average value +/- precision level

Without this information, the statistical significance of a value is indeterminate, meaning that the value itself cannot be assessed.

Examples for considering or neglecting the criteria

Reproducibility

Positive example of reproducibility:

The theory of relativity has been the subject of constant intensive review since the beginning of the 20th century. Many of its predictions could only be investigated and confirmed decades after it was published. Even though its validity is undisputed, it still has not been successfully integrated into the remaining areas of fundamental physics (quantum physics, grand unification theory).
This is one of the best examples of reproducibility. All experimental findings have been confirmed independently multiple times, often repeatedly.

Negative example of reproducibility:

Quote from Nickel prolonged contact, definition by ECHA, 02/04./2014:
"3) *Information from time-related Ni release from alloys including coins (Julander et al., 2009; Lidén et al., 2008; Karolinska Institutet/Lidén C et al., 2013 unpublished opinion).*
4) The amount and rate of Ni (NiCl2 vs NiSO4) permeation/absorption by the skin (Fullerton et al., 1986, Hostynek 2003) including information on skin area permeation/sensitivity (Hostynek 2003)."

The corresponding documents were not made available following an inquiry even three years after publication of the document.
In this instance, supposedly scientific findings consciously and purposefully elude independent testing.

Informative value

Positive example of informative value:

Better Regulation Guidelines of the European Commission, Chapter VI.
https://ec.europa.eu/info/sites/info/files/better-regulation-guidelines-evaluation-fitness-checks.pdf

This document contains clear criteria for assessments. It refers specifically to five criteria (page 52) whose content requirements are specified. They are repeated in roadmaps (e.g. http://ec.europa.eu/smart-regulation/roadmaps/docs/2017_env_005_reach_refit_en.pdf) and include questions.
This is clearly an attempt to provide REFIT studies with uniform empirical significance.

Negative examples of informative value:

1. ESTABLISHING A REFERENCE DOSE RESPONSE RELATIONSHIP FOR CARCINOGENICITY OF HEXAVALENT CHROMIUM, RAC/27/2013/06 Rev. 1

On pages 4 and 5 of the document, a linear "dose-response relationship" is defined based on various studies and subsequent meta-studies.
1. It was not taken into consideration that the vast majority of studies dealt with dust from the production of chromium trioxide compounds. Experience shows that the effect of exposure

depends on physical form. However, the dose-response relationship is currently valid for all physical states (e.g., solid, liquid, aerosol).
2. The exposures were generally greater than 10µg/m^3, whereas the dose-response relationship was attributed validity up to the nanogram range (i.e. 1,000 times lower).
3. Despite relatively short investigation periods, the alleged level of risk was calculated for an 89-year life expectancy; however, exposure at the workplace (and only such studies were available) does not apply until at least 15-20 years of age. Thus, the indicated risk was only estimated for persons with a life expectancy of 100-120 years; otherwise, the "excess cancer risk" in comparison to the background risk would not be correctly determined.
https://echa.europa.eu/documents/10162/13579/rac_carcinogenicity_dose_response_crvi_en.pdf/facc881f-cf3e-40ac-8339-c9d9c1832c32
Obviously, the statements used and the resulting deductions and definitions are not applicable to the original problem and are also outdated. The informative value is thereby invalidated for the assessment of risks in electroplating.

2. Incorrect selection of a database in the case of chromium trioxide

In connection with the allegedly high risk from chromium trioxide use in electroplating, we often use a (meta) study by Seidler which was supported by the Federal Institute for Occupational Safety and Health (BAuA)[1]. Inter alia it forms the basis for the derivation of the currently valid dose-response curve.
This scientific article classified various studies as methodologically inadequate. In the end, only two studies on the relationship between chromium VI concentration in the air and additional cancer risk remained. Here, it was not taken into account that these studies were made in the field of chromium oxide production. Thus, the main polluting element was actually a dust-like substance, whereas only aerosols occur in electroplating. The possible negative effect of dust in its own right was ignored. The meta-study therefore dealt with data that did not concern the scope of galvanic manufacturing and was therefore inapplicable. The study itself was not wrong – but it did not justify all the conclusions and measures that were derived.

Representativeness

Positive example of representativeness:

It is hardly possible to give a positive example, since it would require all effective parameters to be known. Falsification (see negative example) is simple, since only one conflicting example suffices to reject the basic assumptions. Verification requires more detailed and careful individual consideration; however, it should be considered rather as a hypothesis.
Every good study should therefore be comprehensive and reproducible in how it deals with its boundaries, clearly processing and presenting them. Below is a description of the aspects[3].

Methodical definition of representativeness:
A sample is representative of the size of the population to be estimated if and only if the corresponding sample estimate is unbiased.

Unbiased means that the estimate garnered from the sample (e.g. the average income of Germans over the age of 18) does not systematically deviate from the actual value of the population and is

[3] https://www.marktforschung.de/hintergruende/themendossiers/repraesentativitaet-2012/dossier/repraesentativitaet-von-stichproben/

therefore undistorted. With income, distortion would arise for instance if wealthy people are more likely to refuse to participate in a survey than less wealthy individuals, as this would cause the average income to be systematically underestimated. Such a systematic over- or underestimation is called bias. Bias is defined as the difference between the expected value of the estimate (i.e. the mean estimate for theoretically infinite repetition of the sample) and the actual value in the population. Bias usually occurs whenever the sample is distorted in a way that can neither be quantified nor corrected. The main causes of bias are:

- Refusal to participate
- Unknown or uncorrected heterogeneity of selection probabilities (e.g. people who travel are more difficult to reach)
- Unavailability of parts of the population (e.g. households with no landline in telephone samples, address bases with undercoverage in company samples)
- Characteristics of field work and the survey method
- Self-selection
- etc.

It is usually not possible to determine the extent of the bias. And it cannot be reduced by increasing the sample size. Sampling error – i.e. the random deviation of the estimate from the expected value of the estimate – on the other hand does decrease by increasing the sample size. The overall error of the sample is therefore the sum of the bias and the sampling error.

Negative example of representativeness:

Selection of subjects in studies to determine and define prolonged contact for metallic alloys containing nickel

Entry 27 of Annex XVII to REACH refers to the restriction of skin contact to nickel for the general population. No restriction to certain population groups was provided.

ECHA uses a meta-study to quantify the definition of “prolonged contact”. The cited original studies, however, only refer to results with already sensitised subjects. The estimated period after which re-sensitisation or a corresponding physical reaction can be expected is therefore clearly underestimated for the general population. According to the description above, the problems of self-selection and bias were ignored in this instance.

Measures resulting from these assumptions therefore lead to systematic overregulation as the considered population is not representative of the general population.

Accuracy

Positive example of correctness:

Refutation of “cold fusion”: Time and again, this form of energy production appears in articles and achievements are repeatedly reported. But the following quotes from a science magazine bring to light how correctness is tested:

“The two researchers at that time supposedly observed how heavy hydrogen isotopes fused to form helium during electrolysis on the surface of a palladium cathode. Afterwards, however, the two scientists saw themselves as the subjects of massive criticism. Colleagues accused them of gross errors in their experimental procedure, including intentional deception. (...)”

“The best-known method comes from Pons and Fleischmann. The two researchers hypothesised that heavy hydrogen nuclei (deuterium nuclei) can be pulled so strongly into a lattice of palladium atoms that they overcome their repulsive force and fuse together. The two chemists’ observations,

however, could not be confirmed by the majority of researchers who attempted to reproduce the experiments.”
(https://www.spektrum.de/wissen/die-kalte-fusion-wunsch-oder-wirklichkeit/1315962)

Negative example of correctness:

CLP classification of titanium dioxide – by comparison with analogies of detectable systematic errors

Titanium dioxide should be classified as a carcinogen, as studies have shown that it has such an effect.

However, when interpreting the (very few) accessible studies, it was not considered that the carcinogenic effect on mice may well be due to the high intake of dust. This mechanism has been known for a long time.

Since this effect was not considered and the corresponding studies did not carry out comparative studies with chemically different dusts for comparison, it is very likely that a systematic error is present.

Testing the relevance of the thresholds for silica dust, hardwood dust and chromium (VI) compounds – plausibility check

In the document “2016/0130 (COD) – 13/05/2016 Legislative proposal”, the impacts of thresholds for the three types of substance are estimated. The document concludes that all three thresholds have a significant impact on the health of people in Europe. The following data was cited:

a. Respirable silica dust: With the threshold of 0.1 mg/m^3, around 99,000 cancer cases could be prevented by 2069 (= 53 years). This corresponds to a financial saving of 34-89 billion euros.
b. Hardwood dust: 3mg/m^3 should achieve a monetary saving of 12-54 million euros over the same period.
c. Chromium (VI) compounds: Positive effects are expected with the introduction of a threshold of 0.025mg/m^3.

If one checks the relevance of these thresholds for plausibility. the following results are produced:
With an average of 61.5 billion euros for 99,000 cancer cases due to silica dust by 2069, the value of a single case is around 620,000 euros. This means that 1,868 cases would be prevented each year.
With a mean value of 33 million euros for cancer cases due to hardwood dust, €33 million / €620,000 results in a total of 53 cancer cases and thus just a single (statistical) case per year!
For chromium (VI) compounds, it is not possible to make an analogous estimate, because the EU Commission could not provide any values.
In comparison to the 3.7 million new cancer cases that occur annually in Europe (http://www.euro.who.int/en/health-topics/noncommunicable-diseases/cancer/data-and-statistics), there is hardly any relevance for recognising these three thresholds and causes of cancer.
The entire initiative should therefore be rejected because no significant impact on the health of people in Europe can be identified/measured.

Bench test or standard for determining the average fuel consumption of motor vehicles

- ⇨ Tyre dimensions are significantly different from the tyres that are actually used and significantly reduce rolling resistance and moving masses.
- ⇨ Tyre pressure lies above the maximum level, which again reduces consumption compared to actual use.
- ⇨ The test treadmill is chosen so that the friction coefficient is as low as possible.

Since essential parameters systematically deviate from normal usage, the test is not able to quantify the actual conditions. Even comparisons are not possible because it is not possible to transfer the data to actual applications.

Precision

Positive example of precision:

The Hubble constant, i.e. the value that describes the speed at which the universe is expanding, has been measured time and time again since 1920.

Recently, two different redeterminations were made, the results of which differ significantly. Only the high precision of the individual studies made it possible to see this difference. Previously, the spread of the determined values was too high.

According to the scientists, they are now working on seeking out a possible systematic error. Should one be found, the theories would have to be rechecked.

(https://www.spektrum.de/news/expandiert-das-weltall-noch-schneller-oder-nicht/1436550)

Negative example of precision:

For the relationship between concentration in urine and inhalation (respirable) exposure concentration, the following quantitative relationships are provided[4]:

Linear regression	**n**	**r**	**References**
c_U (mol/l) = 0.08 + 0.8 c_A(mg/l)	10	0.95	Rahkonen et al., 1983
c_U (g/g creatinine) = 13.5 + 0.05 c_A(g/m^3)	22	0.52	Raithel, 1987
c_U (g/l) = 10.84 + 0.007 c_A(g/m^3)	174	0.52	Emmerling et al., 1989
c_U (g/l) = 8.49 + 56.88 c_A(mg/m^3)		0.86	Sunderman et al., 1986 *according to data by Morgan and Rogue, 1984

From which the following example values were derived in two tables:

Air Nickel (mg/m^3)	**Urine Nickel (µg/l)**
0.1	15
0.3	30
0.5	45

Air Nickel (mg/m^3)	**Urine Nickel (µg/l) Rahkonen**[5]	**Urine Nickel (µg/l) Raithel**[6]	**Urine Nickel (µg/l) Emmerling**[7]	**Urine Nickel (µg/l) Sundermann**
0.1	9.4	18.5	11.54	15.37
0.3	18.8	28.5	12.94	25.55
0.5	28.2	38.5	14.44	36.93

Conversion factor from mmol/l to mg/l = 58.7 mg/mmol

The calculated individual values show clear differences in both slope and absolute value. A conceivable overlap of the results requires a very broad range of errors that would immediately call into question the individual results.

[4] Proposal by the European Chemical Agency (ECHA) in support of occupational exposure limit values for Nickel and its compounds in the workplace, October 2017, page 47.

[5] The given formula only leads to plausible values if, in contrast to the given mol/l, µmol/l is used as a basis.

[6] The Raithel and Emmerling formulas also only lead to plausible results if the result is assumed to be measured in µg/l or µg/m^3. Apparently, these special characters are missing in the ECHA document. They were inserted by the authors to allow meaningful evaluations.

[7] ditto

Furthermore, the observed exposure values are still far above the usual concentrations found in industry of 1µg/m^3 to 25µg/m^3, i.e. 0.001mg/m^3 to 0.025mg/m^3. With these actual common values, the results are as follows:

Air Nickel (mg/m^3)	**Urine Nickel (µg/l) Rahkonen**	**Urine Nickel (µg/l) Raithel**	**Urine Nickel (µg/l) Emmerling**	**Urine Nickel (µg/l) Sundermann**
0.001	4.74	13.55	10.847	8.55
0.01	5.16	14	10.91	9.06
0.025	5.87	14.75	11.02	9.91

A simple examination of the precision of the individual quantitative data would have shown that the studies used are not suitable as a basis for biomonitoring at the exposure levels encountered in modern production facilities. With these seemingly quantitative correlations, a scientific statement is made that is not actually justified. Over the common exposure range, the results are obviously indistinguishable from the background noise – the intervals of the functions do not overlap. Already the given correlation values and sample sizes would have led to doubts in the event of a serious, solid scientific evaluation and assessment.