PRINCIPLES OF TIERED TEST SYSTEMS

Rudén**, C.** and S.O. Hansson

Royal Institute of technology, Philosophy, Teknikringen 78B, SE-100 44 Stockholm, Sweden

Introduction

Toxicological and ecotoxicological testing of chemicals, aiming at the prediction of adverse effects to human health and the environment, is a resource demanding activity. Producing test data to enable a full risk assessment (i.e. including long-term and multi-generation testing) involves high costs, both in terms of money and in terms of other values such as animal welfare.

In regulatory applications, toxicological tests are combined into test systems. A test system contains rules for when and in what order the different tests should be applied. With the resources presently available it will be necessary to use *tiered systems* in which relatively simple tests are applied to all chemicals that are up for assessment, and the outcomes of these simple tests are used to prioritize substances for further, more resource-intensive testing. Most regulatory test systems currently in use are tiered in the sense that they include the possibility for regulatory agencies to require (additional) testing on the basis of concern raised after initial tests have been performed.[1] A test system thus consists of the individual tests allowed to be used, as well as the rules and criteria that determine which tests are relevant and in what order they should be performed. Although tiered testing has a long tradition, no general theory seems to be available for the combination of single tests into efficient tiered testing systems.

With the implementation of REACH, the proposed new European chemicals legislation, the regulatory division between “new” and “existing” chemicals will be abandoned. As a consequence of this, the forthcoming system will have to handle test requirements for all the 70 000 general chemicals. This has put scientific and regulatory focus on how testing should effectively be performed in the regulatory context.[2, 3] How should chemicals be selected for testing? How extensive testing should be required? What tests should be prioritized?

A new strategy must take into account the limitations in economic resources and testing capacity. It also has to be in line with the aim to reduce the use of animals in toxicological testing.[4,5] Another important aspect is the constant evolution of toxicological knowledge, leading to the identification of previously unknown adverse effects, and subsequently in some cases to a need to develop new predictive tests for these effects. This is currently ongoing for endocrine disruptors.[6,7,8] Such a process may include method development, standardization of test methods, the definition of regulatory test reqirements, criteria, and principles for risk assessment. The combined force of the three abovementioned objectives, (1) to fill data gaps for a large number of chemicals as efficiently as possible, (2) to reduce the use of animals for toxicity testing, and (3) to develop predictive test systems for new endpoints of concern, has put recent focus on the need for developing improved test systems.

Characteristics of single tests

Individual toxicological and ecotoxicological tests can be described in terms of their

(i) cost
(ii) validity
(iii) reliability
(iv) sensitivity

Cost can refer here to the monetary price paid for the execution of a test. Alternatively, the term "cost" can also be used to denote the total social loss or detriment associated with a test. In the latter sense, sacrifice of animal welfare is part of the costs of the tests. By ***validity*** is meant that the test measures what it is intended to measure. The validity of a test thus needs to be evaluated in the light of the purpose of testing. If the purpose is to test whether compound X has a particular adverse effect in a specific strain of mice, then it is of course valid to use this mouse strain for testing. However, if the purpose of testing is to provide information that will

form the basis of human risk assessment, then the experimental species is used to represent another species, namely humans. In this case the validity of a test is difficult to assess, since it needs to include deliberations on potential qualitative and quantitative species differences in metabolism and sensitivity. Such information is usually not readily available. For instance, if we want to investigate the biological effects of exposures to an endocrine disruptor with the aim to make a risk assessment for humans, then to achieve full validity the experimental species must have the same sensitivity to this agent as humans. Since we do not know this, we need at least to ensure that the chosen biological model is *at all* (potentially) sensitive to the exposure under study (this can in some cases be done by using positive controls). The test is, for instance not valid if it is performed on a species (or strain) that lacks the necessary hormone receptor. By ***reliability*** is meant that repeated performance of the test will yield concordant results over time and between laboratories, i.e. that random errors have sufficiently small impact on the outcome. By ***sensitivity*** we mean that the test will identify sufficiently small effects. The sensitivity of a test model is determined by its *statistical power*, i.e. the probability that a study of a given size would detect as statistically significant a real difference of a given magnitude. which in turn is determined by (i) the standard deviation of the exposure, (ii) the standard deviation in the response variable, (iii) the size of the effect that should be measured, (iv) the number of exposed and unexposed subjects included in the study, and (v) the level of statistical significance that is required. How sensitive a test has to be depends, of course, on the regulatory demands. A regulation that aims at avoiding very small effects, such as a small increase in the frequency of a disease, requires a more sensitive test (or test system) than a regulation that only aims at avoiding rather large effects.

It is useful to further summarize how a test can go wrong in terms of the frequencies of the two major types of error. A type I error, or false positive, consists in the test giving an indication of an adverse effect although there is in fact no such effect. A type II error, or false negative, consists in the test giving no indication of an adverse effect although there is in fact such an effect. Although the frequencies of these two types of errors are statistical terms, it is important to realize that they reflect the biological properties of a test system in its relation to the real-life biological system of which it is a model. In practice we selldom have access to the actual frequencies of type I and type II errors for a particular test.

It is important to acknowledge that there is no such thing as the perfect test. If we had, for all important endpoints, tests that fulfill the criteria of low cost, as well as high validity, sensitivity, and reliability, then the scientific uncertainties inherent in testing and risk assessment could be substantially reduced. In reality every test is a trade-off between these aspects.

Combining tests into test systems

Since every test represents a trade-off between (at least some of) the aspects discussed above, we face the challenge to combine tests with different strengths and weaknesses to scientifically well-founded and resource-efficient test systems in which the tests compensate for each other's weaknesses as far as possible. The traditional way of designing test systems is to prioritize low cost at lower tiers (to enable testing of many compounds), whereas the validity of the obtained data is given higher priority at higher tiers (to enable well founded risk management decisions). This implies a strong preference for non-animal models at first tier. Examples of methods used or proposed for first tier testing (priority setting) are *in vitro* models, toxicogenomics, metabolomics, chemical characterization (in particular persistence and bioaccumulation), (Q)SAR parameters, exposure parameters, and group assessment (See e.g. the report "REACH and the need for intelligent testing strategies" (undated) from the Institute for Health and Consumer Protection, available at http://www.jrc.cec.eu.int/download/20051107its.pdf).

It is also part of the standard strategy to choose lower tier test methods so that false negatives (Type II errors) are minimized, while allowing for some false positives (Type I errors). An important reason for this is that the false positives can be corrected at higher tiers, whereas false negatives will not be corrected since they do not reach higher tiers. It should however be noted that the frequency of false positives must not become so high so that priority setting becomes meaningless.

Mechanistic considerations are important in the construction of tiered test systems. In the construction of test systems, results from correlations studies should be combined with mechanistic knowledge. As a general rule, test models that are sensitive to the same mechanism should be arranged serially in the test system (of course taking all other characteristic of the test into account), while mechanistically unrelated test models should preferably be performed in parallel.

A major consideration in the construction of lower tier testing is how to avoid false negatives that may result from lack of sensitivity of the test model. The traditional way to perform tiered testing in ecotoxicology is to start with a complex endpoint (such as survival or reproductive success) in order to cover as many mechanisms as possible. The compounds identified as having an adverse effect in the initial model are then selected for specific testing aiming at identifying the mechanism of action behind the effect. This approach has the disadvantage that the initial test is usually both time and resource consuming (including an in vivo approach and sometimes both long-term and multi-generation exposures). Therefore very few chemicals undergo this type of testing in practice.

An alternative approach is to apply a simple and fast method to many (or all) chemicals at first tier. Some of these simple and fast test models have a limited sensitivity, i.e. they focus on specific endpoints (e.g. receptor or macromolecular binding) and are not valid for other modes of action. If first tier test(s) are invalid for investigating a particular mechanism of action, then chemicals exerting their toxicity by this mechanism will not be identified and thereby potentially excluded from further testing. A promising approach is to combine several tests, each of which has limited validity in this sense (i.e. a narrow mechanistic scope), so that in combination they cover most of the relevant mechanisms. Admittedly, the risk of false negatives may be larger in such combinations than in more complex in vivo tests. However, the construction of first tier testing is always a trade-off with resource limitations. If many more substances can be tested with a much cheaper but somewhat less sensitive test, this may outweigh the loss in validity.

In order to know the predictive value of a simple test, we need to know how results obtained from this test relate to effects on the target system (humans, respectively the ecosystems that the regulation intends to protect). This, however, is seldom known. At best, we can compare the simple test to a more advanced one. This is a far from perfect approximation. Even a state-of-the-art test (such as a long term animal test) provides in its turn only an estimate of effects in humans. The evaluation of first tier methods cannot be more reliable than the advanced tests to which they are compared.

Studies of the correlations between different types of test methods can be used for the validation of tests. [3, 8, 9] Such correlation studies can also be used in the construction of test systems in several other ways. If we consider using a simple test A as a means for priority-setting for a more complex test B, it is important to know how outcomes in the two tests correlate with each other so that we can estimate what we will lose in false positives and false negatives. Furthermore, we want to know if tests used for priority-setting have any selection effects in addition to the intended ones. Hence, tests for persistency and bioaccumulative potential are used for priority-setting for ecotoxicological testing, since we can expect ecotoxicity to have more serious effects in the environment if the substance is also persistent and bioaccumulating. However, it is important to know if tests for persistency and bioaccumulation also tend to select for more toxic or for less toxic substances. In the former case, the usefulness of these tests for priority-setting would increase whereas in the latter case it would decrease.

Several studies have been made of correlations between different endpoints and different test methods. Short-term lethality and data on reproductive performance obtained from life-cycle studies have been shown to correlate to population-level effects[10, 11], and it has also been suggested that population growth (r_m) correlates better to ecosystem risk than does the survival of individuals and reproductive success.[12] Furthermore potential correlations have been investigated between different effect estimates (such as NOAEL, LOAEL, or a benchmark dose) in short-term and long-term tests[13,14,15,16], between different endpoints, such as mutagenicity and carcinogenicity[17, 18, 19, 20, 21], between general (short-term) toxicity and carcinogenicity[22, 23, 24] and between specific short-term toxicity and carcinogenicity.[25] Correlations have also been investigated between chemical properties and toxicity[26, 27], and between effects seen in different species.[28, 29] However, in spite of these individual studies no efforts seem to have been made to study correlations between test results with the purpose of developing better general (regulatory) test strategies feasible for a large number of previously untested chemicals. The development of knowledge in this unexplored field is thus lacking a comprehensive approach, which in our view is necessary for the optimization of testing.

Conclusion

Systematic investigations need to be performed of the relations between different tests, both in terms of toxicological mechanisms and in terms of statistical correlations between their outcomes in different substance groups. Only with major efforts along these lines will it be possible to construct cost-efficient and

reliable test systems that can deal with the major data gaps for general chemicals that was the starting-point of the present investigation.

References

1. Van Leeuwen CJ, Hermens JLM. *Risk assessment of chemicals: An introduction.* Kluwer Academic
2. Combes, R and Balls, M. *ATLA-Alternatives to Laboratory Animals* 2005; 33:289-297.
3. Green S, Goldberg AM, Zurlo J. *Regul Toxicol Pharmacol* 2001;33:105-109.
4. Hareng L, Pellizzer C, Bremer S, et al. *Reproductive Toxicology* 2005;20:441-452.
5. Botham PA. *Toxicology in Vitro* 2004;18:227-230.
6. Carney EW, Hoberman AM, Farmer DR, et al. *Reproductive Toxicology* 1997;11:879-892.
7. Baker VA, *Toxicology in Vitro* 2001;15:413-419.
8. Gelbke HP, Kayser M, Poole A. *Toxicology* 2004;205:17-25.
9. Hushon JM, Clerman RJ, Wagner BO. *Environmental Science & Technology* 1979;13:1202-1207.
Publishers, Dordrecht, NL, 1995.
10. Kuhn A, Munns WR Jr, Champlin D, Mckinney R, Tagliabue M, Serbst J, Gleason T. *Environ Toxicol Chem* 2001;20:213–21.
11. Kuhn A, Munns WR Jr, Poucher S, Champlin D, Lussier S. *Environ Toxicol Chem* 2000;19:2364-2371.
12. Forbes VE, Calow P. *Bioscience* 2002;52:249-257.
13. Weil CS, McCollister DDJ. *Agric Food Chem* 1963;11:486-491.
14. Pieters MN, Kramer HJ, Slob W. *Regul Toxicol Pharmacol* 1998;27:108-111.
15. Kramer HJ, Van den Ham WA, Slob W, Pieters MN. *Regul Toxicol Pharm* 1996;23:249-255.
16. Gaylor DW, Gold LS. *Regul Toxicol Pharmacol* 1998;28:222-225.
17. IARC, 1999. McGregor, Rice and Venitt (Eds.) IARC scientific publications No. 146. International Agency for Research on Cancer.
18. Parodi S, Malacarne D, Romano P, Taningher M. *Environ Health Perspect* 1991;95:199-204.
19. Travis CC, Saulsbury AW, Pack SAR. *Mutagenesis* 1990;5:213-219.
20. McCann J, Gold LS, Horn L, McGill R, Graedel TE, Kaldor J. *Mutat Res* 1988;205:183-195.
21. Piegorsch WW, Hoel DG. *Mutat Res* 1988;96:161-175.
22. Zeise L, Crouch EA, Wilson R. J *Amer Col Toxicol* 1986;5:137-151.
23. Travis CC, Wang LA, Waehner MJ. *Mutagenesis* 1991;6:353-360.
24. Ennever FK, Rosenkranz HS. *Environ Mut* 1986;8:849-865.
25. Albert RE, Magee PS. *Risk Anal* 2000;20:317-325.
26. Benigni R, Andreoli C, Giuliani A J. *Toxicol Environ Health* 1989;27:1-20.
27. Van Haelst AG, Hansen BG. *Environ Toxicol Chem* 2000;19:2372-2377.
28. Gray GM, Li P, Shlyakhter I, Wilson R. *Regul Toxicol Pharmacol* 1995;22:283-291.
29. Bukowski JA, Schnatter AR, Korn L. *Risk Anal* 2001;21:601-611.