"Basic research is like shooting
an arrow in the air and, where it lands,
painting a target."

Homer Adkins, 1984
Nature 312, 212.

Food for Thought
Look Back in Anger – What Clinical Studies Tell Us About Preclinical Work

Thomas Hartung

Johns Hopkins University, Bloomberg School of Public Health, CAAT, Baltimore, USA and University of Konstanz, CAAT-Europe, Germany

Summary

Misled by animal studies and basic research? Whenever we take a closer look at the outcome of clinical trials in a field such as, most recently, stroke or septic shock, we see how limited the value of our preclinical models was. For all indications, 95% of drugs that enter clinical trials do not make it to the market, despite all promise of the (animal) models used to develop them. Drug development has started already to decrease its reliance on animal models: In Europe, for example, despite increasing R&D expenditure, animal use by pharmaceutical companies dropped by more than 25% from 2005 to 2008. In vitro studies are likewise limited: questionable cell authenticity, over-passaging, mycoplasma infections, and lack of differentiation as well as non-homeostatic and non-physiologic culture conditions endanger the relevance of these models. The standards of statistics and reporting often are poor, further impairing reliability. Alarming studies from industry show miserable reproducibility of landmark studies. This paper discusses factors contributing to the lack of reproducibility and relevance of pre-clinical research. The conclusion: Publish less but of better quality and do not rely on the face value of animal studies.

Keywords: preclinical studies, animal studies, in vitro studies, toxicology, safety pharmacology

Introduction

The prime goal of biomedicine is to understand, treat, and prevent diseases. Drug development represents a key goal of research and the pharmaceutical industry. A devastating attrition rate of more than 90% for substances entering clinical trials has received increasing attention. Obviously, we often are not putting our money on the right horses… Side effects not predicted in time from toxicology and safety pharmacology contribute 20-40% to these failures, indicating limitations of the toolbox, which is considerably larger than what is applied to environmental chemicals, with the exception of pesticides. Here, the question is raised whether quality problems of the disease models and basic (especially academic) research also contribute to this. In a simplistic view, clinical trials are based on the pillars of basic research/pre-clinical drug development, and toxicology (Fig.1).

What does this tell us for areas where we have few or no clinical trials to correct false conclusions? Toxicology is a prime example, where regulatory decisions for products traded at $ 10

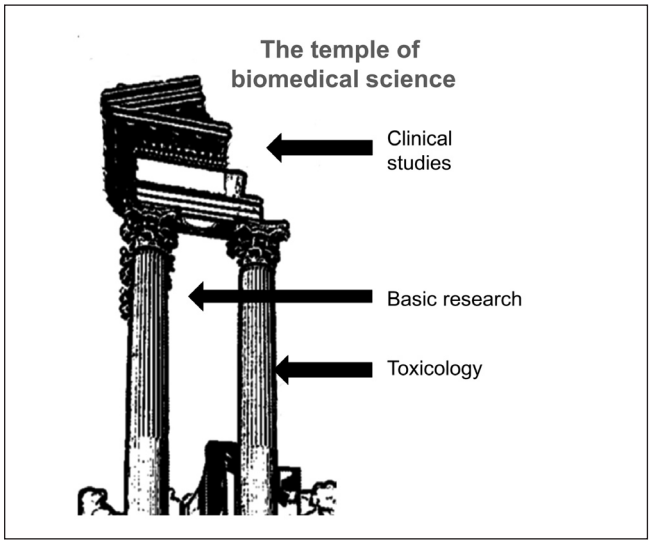

Fig. 1: Clinical trials are based on the pillars of basic research / pre-clinical drug development, and toxicology

trillion per year are taken only on the basis of such testing (Bottini and Hartung, 2009, 2010). Are we sorting out the wrong candidate substances? Aspirin likely would fail the preclinical stage today (Hartung, 2009c). Rats and mice predict each other for complex endpoints with only 60% accuracy and, predicted together, only 43% of clinical toxicities of candidate drugs observed later (Olson et al., 2000). New approaches that rely on molecular pathways of human toxicity currently are emerging under the name "Toxicology for the 21st Century".

Doubt as to animal models also is increasing: A number of increasingly systematic reviews summarized here more and more show the limitations. A National Academy of Sciences panel recently analyzed the suitability of animal models to assess the human efficacy of countermeasures to bioterrorism: It could neither identify suitable models nor did it recommend their development; it did, however, call for the establishment of other human-relevant tools. In line with this, about $ 200 million have been made available by NIH, FDA, and DoD agencies over the last year to start developing a human-on-a-chip approach (Hartung and Zurlo, 2012).

Academic research represents a major stimulus for drug development. Obviously, basic research also is carried out in pharmaceutical industry, but quality standards are different and the lesser degree of publication makes them less accessible for analysis. Obviously, academic research comes in many flavors, and when pinpointing some critical notions here, each and every one might be unfair and not hold for a given laboratory. Similarly, the author and his generations of students are not free from the alleged (mis)behaviors. It is the far too frequent, retrospective view, imprinted from experiences from quality assurance and validation that will be shared here.

Consideration 1: The crisis of drug development

The situation is clear: Companies spend more and more money on drug development, with an average of $ 4 and up to $ 11 billion quoted by Forbes for a successful launch to the market[1]. The number of substances making it to market launch is dropping, and their success does not necessarily compensate for the increased investment. The blockbuster model of drug industry seems largely busted.

The situation was characterized earlier (Hartung and Zurlo, 2012), and more recent figures do not suggest any turn for the better: Failure rates in the clinical phase of development now reach 95% (Arrowsmith, 2012). Analysis by the Centre for Medicines Research (CMR) of projects from a group of 16 companies (representing approximately 60% of global R&D spending) in the CMR International Global R&D database reveals that the Phase II success rates for new development projects have fallen from 28% (2006-2007) to 18% (2008-2009) (Arrowsmith, 2011a). 51% were due to insufficient efficacy, 29% were due to strategic reasons, and 19% were due to clinical or preclinical safety reasons. The average for the combined success rate at Phase III and submission has fallen to ~50% in recent years (Arrowsmith, 2011b). Taken together, clinical phases II & III now eliminate 95% of drug candidates.

This appeared to correspond to dropping numbers of new drugs, as observed between 1997 and 2006, as we have occasionally referenced (Bottini and Hartung, 2009, 2010), though this has been shown to be possibly largely an artifact (Ward et al., 2013). We also have to consider that attrition does not end with the market launch of drugs: Unexpected side effects lead to withdrawals – Wikipedia, who knows it all, lists 47 drugs withdrawn from the market since 1990[2], which represents roughly the number of new drug entities entering the market in two years. This does not even include the drugs for which indications had to be limited because of problems. There also are examples of drugs that made it through the trials to the market but, in retrospect, did not work (see for examples the AP press coverage in October 2009 following the US Government Accountability Office report analyzing 144 studies, and showing that the FDA has never pulled a drug off the market due to a lack of required follow-up about its actual benefits[3]).

At the same time, combining the results of 0.32% fatal adverse drug reactions (ADR) (Lazarou et al., 1998) (total 6.7% ADR) of all hospitalized patients in the US in 1998, with a 2.7-fold increase of fatal ADR from 1998-2005 (Moore et al., 2007), leads to about 1% of hospitalized patients in the US dying from ADR. This suggests that drugs are not very safe, even after all the precautionary tests, and corresponds to the relatively frequent market withdrawals.

The result of this disastrous situation is that pharma companies are eating each other up, often in the hope of acquiring a promising drug pipeline, only to find out that this was wishful thinking or losing so much time in the merger that the delay of development compromises the launch of the pipeline drugs.

Consideration 2: Clinical research, perverted by conflict of interest or role model?

A popular criticism of clinical drug development (as, e.g., prominently stressed in Ben Goldacre's recent book "Bad Pharma", 2012) is the bias from the pressure to get drugs to the market. In fact, there is also a publication bias, i.e., the more successful a clinical study, the more likely it will be published. It has been shown that studies sponsored by industry are seven times more likely to have positive outcomes than those that are investigator-driven (Bekelman et al., 2003; Lexchin, 2003). However, this does not take into account how much more development efforts go into industrial preclinical drug

[1] http://www.forbes.com/sites/matthewherper/2012/02/10/the-truly-staggering-cost-of-inventing-new-drugs/

[2] http://en.wikipedia.org/wiki/List_of_withdrawn_drugs (Accessed June 21, 2013)

[3] http://usatoday30.usatoday.com/news/health/2009-10-26-fda-drugs_N.htm?csp=34

development compared to what academic researchers have at their disposal.

Actually, clinical studies have extremely high quality standards: They are mostly randomized, double-blind, and placebo-controlled, as well as usually multi-centric. They require ethical review, follow Good Clinical Practice, and are carried out by skilled professionals. In recent years, the urge to publish and register has increased strongly. Clinical medicine also brought about Evidence-based Medicine (EBM), which we have several times praised as an objective, transparent, and conscientious way to condense information for a given controversial question (Hoffmann and Hartung, 2006; Hartung, 2009a, 2010). All together, these attributes are difficult to match in other fields.

So we might say that clinical research is pretty good even in acknowledging its biases, if at all, of overestimating success. In a simple view, the clinical pipeline, despite enormous financial pressures, has very sophisticated tools to promote good science. If this is true, we put our money on the wrong horses in clinical research to begin with. We have to analyze the weaknesses of the preclinical phase to understand why we are not improving attrition rates.

Consideration 3: Bashing animal toxicology again?

Sure, to some extent. It is one purpose of this series of articles to collect arguments for transitioning to new tools. The quoted data from Arrowsmith would suggest that toxic side-effects contribute to 20% of attrition each in phase II and III. Probably, we need to add some percent for side-effects noted in phase I, i.e., first in humans, and post-market adverse reactions. Thus an overall figure of 30-40% seems realistic.

However, we first have to distinguish two matters: One is the observed effects in humans, which were not sufficiently anticipated. Another is the findings in animal toxicity studies done in parallel to the clinical studies. It is a common misunderstanding among lay audiences that clinical studies commence after toxicology has been completed. For reasons of timing, however, this is not possible, and the long-lasting studies are done at least in parallel to phase II. Currently, when first acquiring data on humans, animal toxicology is incomplete. The two types of toxicological data also are very different: The toxicological effects observed in human trials of necessarily short duration and little or no follow-up observation are necessarily different from the chronic systemic animal studies at higher doses. Fortunately, typical side-effects in clinical trials are mild, the most common one (about half of the cases) is drug-induced liver injury (DILI), observed as a painless and normally easily reversible increase in liver enzymes in blood work though possibly extending to the more severe and life-threatening liver failure. The Innovative Medicine Initiative has tackled this problem in a project based on an initiative we started with industry at ECVAM: "*Many medicines are harmful to the liver, and drug-induced liver injury (DILI) now ranks as the leading cause of liver failure and transplantation in western countries. However, predicting which drugs will prove toxic to the liver is extremely difficult, and often problems are not detected until a drug is already on the market*."[4] The hallmark paper by Olson et al. (2000) gives us some idea of this and the retrospective value of animal models in identifying such problems: "*Liver toxicity was only the fourth most frequent HT [human toxicity]…, yet it led to the second highest termination rate. There was also less concordance between animal and human toxicity with regard to liver function, despite liver toxicity being common in such studies. There was no relation between liver HTs and therapeutic class*."

A completely different question is: What animal findings obtained parallel to clinical trials lead to abandoning substances? Probably not that many. Cancer studies are notoriously false positive (Basketter et al., 2012), even for almost half of the tested drugs on the market; furthermore, genotoxicants usually have been excluded earlier. Reproductive toxicity will lead mainly to a warning against using the substance in pregnancy, which is a default for any new drug, as nobody dares to test on pregnant women. The acute and topical toxicities have been evaluated before being applied to humans. The same holds true for safety pharmacology, i.e., the assessment of cardiovascular, respiratory, and neurobehavioral effects, as well as excess target pharmacology. This leaves us with organ toxicities in chronic studies. In fact, if not sorted out by "investigative" toxicology, this can impede or delay drug development. "Fortunately," different animal species often do not agree as to the organ of toxicity manifestation, leaving open a lot of room for discussion as to translation to humans.

Compared to clinical studies, toxicology has some advantages and some disadvantages as to quality: First, there are internationally harmonized protocols (especially ICH and OECD) and Good Laboratory Practice to quality-assure their execution. However, we use outdated methods, mainly introduced before 1970, which were systematically rendered precautionary/oversensitive, e.g., by using extremely high doses. The mechanistic thinking of a modern toxicology comes as "mustard after the meal," mainly to argue why the findings are not relevant to humans. What is most evident when comparing approaches: clinical studies have one endpoint, good statistics, and hundreds to thousands of treated individuals with relevant exposures. Toxicology does just the opposite: Group sizes of identical twins (inbred strains) are minimal, and we study a large array of endpoints at often "maximum tolerated doses" without proper statistics. The only reason is feasibility, but these compromises combine in the end to determine the relevance of the prediction made. We have made these points in more detail earlier (Hartung, 2008a, 2009b). For a somewhat different presentation, please see Table 1 which combines

[4] http://www.imi.europa.eu/content/mip-dili

Tab. 1: Differences between and methodological problems of animal and human studies critical to prediction of substance effects

Subjects
- Small groups of (often inbred, homogenous genetic background) animals vs. large groups of individuals with heterogeneous genetic background
- Young adult animals vs. all ages in human trials
- Animals typically only of one gender
- Disparate animal species and strains, with a variety of metabolic pathways and drug metabolites, leading to variation in efficacy and toxicity

Disease models
- Artificial diseases, i.e., different models for inducing illness in healthy animals or injury with varying similarity to the human condition of sick people
- Acute animal models for chronic phenomena
- Monofactorial disease models vs. multifactorial ones in humans
- Especially in knock-out mouse models the adaptive responses in animals are underestimated compensating for the knock-out

Doses
- Variations in drug dosing schedules (therapeutic to toxic) and regimens (usually once daily) that are of uncertain relevance to the human condition (therapeutic optimum)
- Pharmaco- and toxicokinetics of substances differ between animals and humans

Circumstances
- Uniform, optimal housing and nutrition vs. variable human situations
- Animals are stressed
- Never concomitant therapy vs. frequent ones in humans

Diagnostic procedures
- No vs. intense verbal contact
- Limited vs. extensive physical exam in humans
- Limited standardized vs. individualized clinical laboratory examination in humans
- Predetermined timing vs. individualized one in humans
- Extensive histopathology vs. exceptional one in humans
- Length of follow up before determination of disease outcome varies and may not correspond to disease latency in humans
- Especially in toxicological studies the prevalence of health effects is rarely considered when interpreting data

Study design
- Variability in the way animals are selected for study, methods of randomization, choice of comparison therapy (none, placebo, vehicle), and reporting of loss to follow up
- Small experimental groups with inadequate power, simplistic statistical analysis that does not account for potential confounding, and failure to follow intention to treat principles
- Nuances in laboratory technique that may influence results may be neither recognized nor reported, e.g., methods for blinding investigators
- Selection of a variety of outcome measures, which may be disease surrogates or precursors and which are of uncertain relevance to the human clinical condition
- Traditional designs, especially of guideline studies, offering standardization but prohibiting progress

The table combines arguments from (Olson et al., 2000; Pound et al., 2004; and Hartung, 2008a).

arguments from different sources (Pound et al., 2004; Olson et al., 2000; Hartung, 2008a) showing reasons for differences between animal studies and human trials.

Consideration 4: Sorting out substances with precautionary toxicology before clinical studies? The case of genotoxicity assays

Perhaps the even more important question with regard to attrition is, which substances never make it to clinical trials, that would have succeeded but whose progress was hindered by wrong or precautionary toxicology? Again we have to ask, what findings lead to the abandonment of a substance. This is more complicated than it seems, because it depends on when in the development process such findings are obtained and what the indication of the drug is. To put it simply, a new chemotherapy will not be affected very much by any toxicological finding. In early screening, we tend to be generous in excluding substances that appear to have liabilities. An interesting case here is genotoxicity – due to the fear of contributing to cancer and the difficulty of identifying human carcinogens at all, this often is a brick wall. In addition, the relatively easy and cheap as-

sessment of genotoxicity with a few *in vitro* tests allows front-loading of such tests. Typically, substances will be sorted out if found positive. The 2005 publication of Kirkland et al. gave the stunning result that while the combination of three genotoxicity tests achieves a reasonable sensitivity of 90+% for rat carcinogens, also more than 90% of non-carcinogens are false positive, i.e., a miserable specificity. Among the false positives are common table salt and sugar (Pottenger et al., 2007). With such a high false positive rate, we would eliminate an incredibly large part of the chemical universe at this stage.

This view has been largely adapted, leading to an ECVAM workshop (Kirkland et al., 2007) and follow-up work (Lorge et al., 2008; Fellows et al., 2008; Pfuhler et al., 2009, 2010; Kirkland, 2010a,b; Fowler et al., 2012a,b) financed by Cosmetics Europe and ECVAM, and finally changes in the International Conference on Harmonization (ICH) guidance, though not yet at the OECD, which did not go along with the suggested 10-fold reduction in test dose for the mammalian assays.

However, the "false positive" genotoxicity issue (Mouse Lymphoma assay and Chromosomal Aberration assay) has been challenged more recently. Gollapudi et al. from Dow presented an analysis of the Mouse Lymphoma Assay at SOT 2012. "*Since the MLA has undergone significant procedural enhancements in recent years, a project was undertaken to reevaluate the NTP data according to the current standards (IWGT) to assess the assay performance capabilities. Data from more than 1900 experiments representing 342 chemicals were examined against acceptance criteria for background mutant frequency, cloning efficiency, positive control values, and appropriate dose selection. In this reanalysis, only 17% of the experiments and 40% of the "positive" calls met the current acceptance standards. Approximately 20% of the test chemicals required >1000 ug /mL to satisfy the criteria for the selection of the top concentration. When the concentration is expressed in molarity, approximately 58, 32, and 10% of the chemicals required ≤1 mM, >1 to ≤10 mM, and >10 mM, respectively, to meet the criteria for the top concentration. More than 60% of the chemicals were judged as having insufficient data to classify them as positive, negative, or equivocal. Of the 265 chemicals from this list evaluated by Kirkland et al. (2005, Mutat Res., 584, 1), there was agreement between Kirkland calls and our calls for 32% of the chemicals*."

Astra-Zeneca (Fellows et al., 2011) published their most recent assessment of 355 drugs and found 5% unexplained positives in the Mouse Lymphoma Assay: "*Of the 355 compounds tested, only 52 (15%) gave positive results so, even if it is assumed that all of these are non-carcinogens, the incidence of 'false positive' predictions of carcinogenicity is much lower than the 61% apparent from analysis of the literature. Furthermore, only 19 compounds (5%) were positive by a mechanism that could not be associated with the compounds primary pharmacological activity or positive responses in other genotoxicity assays*."

Snyder and Green (2001) earlier found less dramatic false positive rates for marketed drugs. FDA CDER did a survey on the most recent ~750 drugs and found that positive mammalian genotoxicity results (CA or MLA) did not affect drug approval substantially (Dr Rosalie Elesprue, personal communication). Only 1% was put on hold for this cause. However, this obviously addresses a much later stage of drug development, at which most genotoxic substances already have been excluded.

In contrast, an analysis by Dr Peter Kasper of nearly 600 pharmaceuticals submitted to the German medicines authority (BfArM) between 1995 and 2005, gave 25-36% positive results in one or more mammalian cell tests, and yet few were carcinogenic (Blakey et al., 2008). It is worth noting that an evaluation by the Scientific Committee on Consumer Products (SCCP) of genotoxicity/mutagenicity testing of cosmetic ingredients without animal experiments[5] showed that 24 hair dyes tested positive *in vitro* were all then found negative *in vivo*. This would be very much in line with the Kirkland et al. analysis. However, we argued earlier (Hartung, 2008b): "*The question might, however, be raised whether mutagenicity in human cells should be ruled out at all by an animal test. A genotoxic effect in vitro shows that the substance has a property, which could be hazardous. Differences in the in vivo test can be either species-specific (rat versus human) or due to kinetics (does not reach the tissue at sufficiently high concentrations). These do not necessarily rule out a hazard toward humans, especially in chronic situations or hypersensitive individuals. This means that the animal experiment may possibly hide a hazard for humans*."

In conclusion, flaws in the current genotoxicity test battery are obvious. There is promise of new methods, most obviously of the micronucleus test, which was formally validated and led to an OECD test guideline. There is some validation for the COMET assay (Ersson et al., 2013), which compared 27 samples in 14 laboratories using their own protocols; the variance observed was mainly between laboratories/protocols, i.e., 79%. Thus standardization of the COMET assay is essential, and we are desperately awaiting the results of the Japanese validation study for the COMET assay *in vivo* and *in vitro*. New assays based, e.g., on DNA repair measurement promise better accuracy (e.g., Walmsley, 2008; Moreno-Villanueva et al., 2009, 2011). Whether the current data justify eliminating the standard *in vitro* tests and adopting the *in vivo* comet assay as specified in the new ICH S2 guidance before validation can be debated. This guidance in fact decreases *in vitro* testing and increases *in vivo* testing (in its option 2 as it replaces *in vitro* mammalian tests entirely with two *in vivo* tests). It is claimed that they can be done within ongoing sub-chronic testing, but this still needs to be shown because the animal genotoxicity tests require a short term (2-3 day) high dose, while the sub-chronic testing necessitates lower doses.

What to do? We need an objective assessment of the evidence concerning the reality of "false positives." This could be a very promising topic for an evidence-based toxicology collaboration (EBTC[6]) working group. Better still, we should try to find a better way to assess human cancer risk without animal testing. The animal tests are not sufficiently informative.

[5] http://ec.europa.eu/health/ph_risk/committees/04_sccp/docs/sccp_s_08.pdf

[6] http://www.ebtox.com

What does this mean in the context of the discussion here? It shows that even the most advanced use of *in vitro* assays to guide drug development is not really satisfactory. Though the extent of false positives, i.e., innocent substances not likely to be developed further to become drugs, is under debate, it appears that no definitive tool for such decisions is available. The respective animal experiment does not offer a solution to the problem, as it appears to lack sensitivity. Thus, the question remains whether genotoxicity as currently applied guides our drug development well enough.

Consideration 5: If animals were fortune tellers of drug efficacy, they would not make a lot of money...

A large part of biomedical research relies on animals. John Ioannidis recently showed that almost a quarter of the articles in PubMed show up with the search term "animal," even a little more than with "patient" (Ioannidis, 2012). While there is increasing acknowledgement that animal tests have severe limitations for toxicity assessments, we do not see the same level of awareness for disease models. The hype about genetically modified animal models has fueled this naïve appreciation of the value of animal models.

The author had the privilege to serve on the National Academy of Science panel on animal models for countermeasures to bioterrorism. We have discussed this recently (Hartung and Zurlo, 2012): The problem for developing and stockpiling drugs for the event of biological/chemical terrorism or warfare is that (fortunately) there are no patients to test on. So, the question to the panel was how to substitute in line with the animal rule of FDA with suitable animal models. In a nutshell, our answer is: There are no such things as sufficiently predictive animal models to substitute for clinical trials (NRC, 2011). Any drug company would long to have such models for drug development, as the bulk of development costs is incurred in the clinical phase; for countermeasures we have the even more difficult situation of unknown pathophysiology, limitations to experiment in biosafety facilities, disease agents potentially designed to resist interventions, and mostly peracute diseases to start with. So an important part of the committee's discussions dealt with the attrition (failure) rate of drugs entering clinical trials (see above), which does not encourage using animal models to substitute for clinical trials at all.

In line with this, a recent paper by Seok et al. (2013) showed the lack of correspondence of mouse and human responses in sepsis, probably the clinical condition closest to biological warfare and terrorism. We discussed this earlier (Leist and Hartung, 2013) and here only one point shall be repeated, i.e., though not necessarily as prominent and extensive, several assessments of animal models led to disappointing results, as referenced in the comment for stroke research.

In toxicology, we have seen that different laboratory species exposed to the same high doses predict each other no better than 60% – and there is no reason to assume that any of them predict humans better at low doses. We lack such analysis for drug efficacy models systematically comparing outcomes in different strains or species of laboratory animals. It is unlikely that results are much better.

In this series (Hartung, 2008a) we have addressed the shortcomings of animal tests in general terms. Since then, the weaknesses in quality and reporting of animal studies, especially, have been demonstrated (MacCallum, 2010; Macleod and van der Worp, 2010; Kilkenny et al., 2010; van der Worp and Macleod, 2011), further undermining their value. Randomization and blinding rarely are reported, which can have important implications, as it has been shown that animal experiments carried out without either are five times more likely to report a positive treatment effect (Bebarta et al., 2003). Baker et al. (2012) recently gave an illustration of poor reporting on animal experiments, stating that in "*180 papers on multiple sclerosis listed on PubMed in the past 6 months, we found that only 40% used appropriate statistics to compare the effects of gene-knockout or treatment. Appropriate statistics were applied in only 4% of neuroimmunological studies published in the past two years in Nature Publishing Group journals, Science and Cell*" (Baker et al., 2012).

Some more systematic reviews of the predictive value of animal models have been little favorable, see Table 2 (Roberts, 2002; Pound et al., 2004; Hackam and Redelmeier, 2006; Perel et al., 2007; Hackam, 2007; van der Worp et al., 2010). Hackman and Redelmeier (Hackam and Redelmeier, 2006), for example, found that of 76 highly cited animal studies, 28 (37%; 95% confidence interval [CI], 26%-48%) were replicated in human randomized trials, 14 (18%) were contradicted by randomized trials, and 34 (45%) remain untested. This is actually not too bad, but the bias to highly cited studies (range 639 to 2233) already indicates that these studies survived later repetitions and translation to humans. There are now even more or less "systematic" reviews of the systematic reviews (Pound et al., 2004; Mignini and Khan, 2006; Knight, 2007; Briel et al., 2013), showing that there is room for improvement. They definitely do not have the standard of evidence-based medicine. In the context of evidence-based medicine, "*A systematic review involves the application of scientific strategies, in ways that limit bias, to the assembly, critical appraisal, and synthesis of all relevant studies that address a specific clinical question*" (Cook et al., 1997). But the concept is maturing. See, for example, the NC3R whitepaper "*Systematic reviews of animal research*"[7] or the "*Montréal Declaration on Systematic Reviews of Animal Studies*."[8] The ARRIVE guideline (Kilkenny et al., 2010) and the Gold Standard Publication Checklist (GSPC) to improve the quality of animal studies (Hooijmans et al., 2010) facilitate the evaluation and standardization of publications on animal studies.

No wonder that *in vitro* studies are increasingly considered: "According to a new market report by Transparency Market Research, the global *in vitro* toxicity testing market was worth $1,518.7 million in 2011 and is expected to reach $4,114.1 million in 2018, growing at a CAGR of 15.3 percent from 2013 to

[7] http://www.nc3rs.org.uk/downloaddoc.asp?id=695

[8] http://www.wc8.ccac.ca/files/WC8_Declaration_of_Montreal_FINAL.pdf

Tab. 2: Examples of more systematic evaluations of the quality of animal studies of drug efficacy

First author	Year published	(Number of) indications	Number of studies considered (of total)	Reproducible in humans
Horn	**2001**	**stroke**	**20 (225)**	**50%**
The methodological quality of the animal studies was found to be poor. Of the included studies, 50% were in favor of nimodipine (which was not effective in human trials). In-depth analyses showed statistically significant effects in favor of treatment (10 studies) (Horn et al., 2001).				
Corpet	**2003**	**dietary agents on colorectal cancer**	**111**	**55%**
"*We found that the effect of most of the agents tested was consistent across the animal and clinical models*." Data extracted from Table 3 (Corpet et al., 2003) with noted discrepant results for 20 studies, but only summary results provided. No quality assurance of data or inclusion/exclusion criteria. Human study end point is not cancer incidence but adenoma recurrence. The two animal models in rat and mice showed a significant correlation of agents tested in both models (r = 0.66; n = 36; P < 0.001). Updated very similar analysis published (Corpet et al., 2005).				
Perel	**2007**	**diverse (6)**	**230**	**50% (of indications)**
"*Discordance between animal and human studies may be due to bias or to the failure of animal models to mimic clinical disease adequately*." Poor quality of animal studies noted.				
Bebarta	**2003**	**emergency medicine**	**290**	**n.a.**
"*Animal studies that do not utilize RND [randomization] and BLD [blinding] are more likely to report a difference between study groups than studies that employ these methods*" (Bebarta et al., 2003).				
Pound	**2004**	**diverse (6)**	**n.a.**	**n.a.**
Analysis of 25 systematic reviews on animal studies found; summary of six examples (Horn et al., 2001; Lucas et al., 2002; Roberts et al., 2002; Mapstone et al., 2003; Ciccone and Candelise, unpublished; Petticrew and Davey Smith, 2003). "*Much animal research into potential treatments for humans is wasted because it is poorly conducted and not evaluated through systematic reviews*."				
Sena	**2010**	**stroke**	**1359**	**n.a.**
Analysis of 16 systematic reviews of interventions tested in animal studies of acute ischemic stroke involving 525 unique publications. Publication bias was highly prevalent (Sena et al., 2010).				
Hackam	**2006**	**diverse**	**76**	**37%**
"*Only about a third of highly cited animal research translated at the level of human randomized trials*." (Hackam and Redelmeier, 2006)				

2018."[9] Compare this to our estimate of $ 3 billion for *in vivo* toxicology (Bottini and Hartung, 2009). The quality problem, however, is no less for *in vitro*: Our attempts to establish Good Cell Culture Practice (GCCP; Coecke et al., 2005) and publication guidance for *in vitro* studies (Leist et al., 2010) desperately await broader implementation (see below).

Consideration 6: Basic research as the start of drug development

Two recent publications by authors from two major pharmaceutical companies provided an epiphany: Both Amgen and Bayer HealthCare showed that they essentially could not reproduce the key findings of many studies that had prompted drug development. Prinz et al. (2011) from Bayer HealthCare stated in Nature Reviews in Drug Discovery "*Believe it or not: how much can we rely on published data on potential drug targets? …data from 67 projects, … This analysis revealed that only in ~20-25% of the projects were the relevant published data completely in line with our in-house findings... In almost two-thirds of the projects, there were inconsistencies between published data and in-house data that either considerably prolonged the duration of the target validation process or, in most cases, resulted in termination of the projects*."

Similarly, Begley and Ellis (2012) from Amgen in Nature "*Raise standards for preclinical cancer research … Fifty-three papers were deemed 'landmark' studies …scientific findings*

[9] Global in-vitro toxicity testing market to take off as push towards alternatives grows By Michelle Yeomans, 05-Nov-2012. http://bit.ly/12OlQs8.ly/12OlQs8

were confirmed in only 6 (11%) cases. Even knowing the limitations of preclinical research, this was a shocking result."

How is this possible? Basic researchers seem to be even more naïve in the interpretation of their results than clinical researchers. In a comparison of 108 studies (Lumbreras et al., 2009), laboratory scientists were 19-fold more likely to over-interpret the clinical utility of molecular diagnostic tests than clinical ones. Basic research, at least in academia, the source of most of such papers, is done mostly unblinded in a single laboratory. It is executed by students learning on the job, normally without any formal quality assurance scheme. Limited replicates due to limited resources and time as well as pressure to publish lead to publications, which do not always stand replication. Insufficient documentation aggravates the situation.

Figure 2 shows a cartoon of some of the problems. Having supervised some 50 PhD and a similar number of master and bachelor students, the author is not innocent of any of these misdoings.

The problem starts with setting the topic; this is rarely as precise as in drug development: Often it simply continues work of a previous student, who left uncompleted work behind after finishing a degree. In other cases it starts as pure exploration with the idea to go into a new direction. How often have we had to change topics or circumstances led us to take up new directions? Still, there is a desire to make use of the work done so far. It is always appealing to combine, reshuffle, etc. in order to make best use of the pieces. The quality of the pieces? Let's be honest: "A typical result out of three" usually means "the best I have achieved." Especially critical is outlier removal: even if following a certain formal process, this is hardly ever properly documented. If things are not significant, we add more experiments, happily ignoring that this messes up the significance testing. Replications are a problem in themselves. How often are these just technical replicates, i.e., parallel experiments and not real reproductions on another day? If the reviewer is not very picky this will fly far too often. Who then combines the different independent experiments with an appropriate error propagation taking into account the variance of each reproduction? Even among seasoned researchers, I have met few who know how to do this.

Using spreadsheets and other interactive data manipulation and analysis tools we do not provide a usable audit trail of how

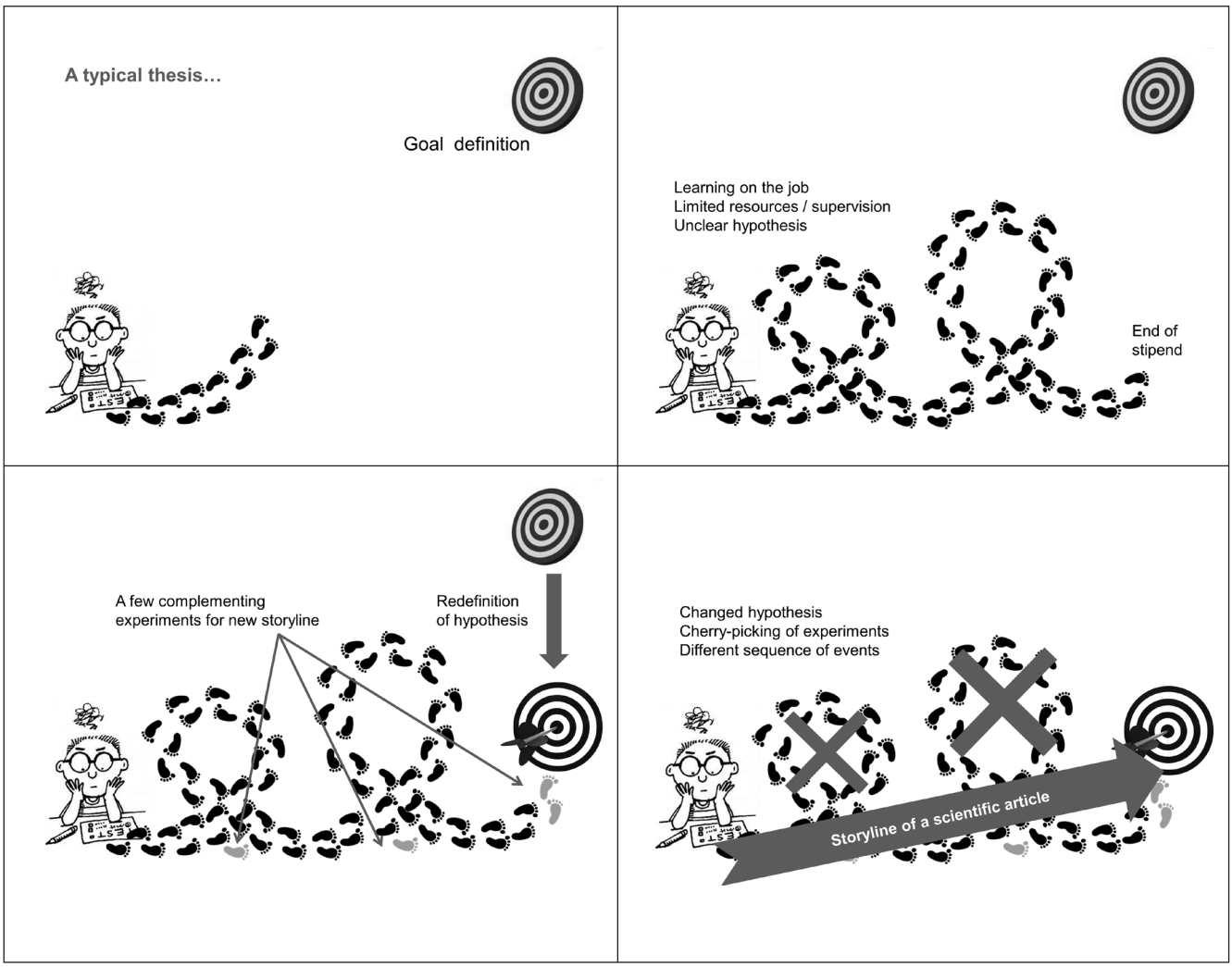

Fig. 2: Typical problems commonly causing overinterpretation of results in basic research

Tab. 3: Twenty Statistical Errors Even YOU Can Find in Biomedical Research Articles
(reproduced with permission of the *Croat Med J* from Lang (2004))

#1:	Reporting measurements with unnecessary precision
#2:	Dividing continuous data into ordinal categories without explaining why or how
#3:	Reporting group means for paired data without reporting within-pair changes
#4:	Using descriptive statistics incorrectly
#5:	Using the standard error of the mean (SEM) as a descriptive statistic or as a measure of precision for an estimate
#6:	Reporting only P values for results
#7:	Not confirming that the data met the assumptions of the statistical tests used to analyze them
#8:	Using linear regression analysis without establishing that the relationship is, in fact, linear
#9:	Not accounting for all data and all patients
#10:	Not reporting whether or how adjustments were made for multiple hypothesis tests
#11:	Unnecessarily reporting baseline statistical comparisons in randomized trials
#12:	Not defining "normal" or "abnormal" when reporting diagnostic test results
#13:	Not explaining how uncertain (equivocal) diagnostic test results were treated when calculating the test's characteristics (such as sensitivity and specificity)
#14:	Using figures and tables only to "store" data, rather than to assist readers
#15:	Using a chart or graph in which the visual message does not support the message of the data on which it is based
#16:	Confusing the "units of observation" when reporting and interpreting results
#17:	Interpreting studies with nonsignificant results and low statistical power as "negative," when they are, in fact, inconclusive
#18:	Not distinguishing between "pragmatic" (effectiveness) and "explanatory" (efficacy) studies when designing and interpreting biomedical research
#19:	Not reporting results in clinically useful units
#20:	Confusing statistical significance with clinical importance

results were obtained and how many attempts were made until significant results were obtained (Harrell, 2011). Poor statistics are a more widespread problem than outsiders might believe. They are a core part of the "*Follies and Fallacies in Medicine*" (Skrabanek and McCormick, 1990). Des McHale coined it: "*The average human has one breast and one testicle*." Awareness is a little better in clinical research (Andersen, 1990; Altman, 1994, 2002), but as reviewers or readers we too often see papers without statistics or with inappropriate statistics (such as the promiscuous use of t-tests where not justified). Some common mistakes were illustrated in (Festing, 2003; Lang, 2004; Altman, 1998) (see also Tab. 3).

Douglas Altman (Altman, 1998) summarized in 1998 thirteen previous analyses of the quality of statistics in medical journals (Tab. 4). The 1667 papers analyzed show that only about 37% have acceptable statistics. No trend to the better is visible.

An example from environmental chemistry is the most commonly used method to deal with values below detection limits, which is to substitute a fraction of the detection limit for each non-detect (Helsel, 2006): "*Two decades of research has shown that this fabrication of values produces poor estimates of statistics, and commonly obscures patterns and trends in the data. Papers using substitution may conclude that significant differences, correlations, and regression relationships do not exist, when in fact they do. The reverse may also be true*."

Tab. 4: Summary of some reviews of the quality of statistics in medical journals, showing the percentage of "acceptable" papers (of those using statistics)

First author and year	Number of papers (journals)	% Papers acceptable
Schor, 1966	295 (10)	28
Gore, 1977	77 (1)	48
White, 1979	139 (1)	55
Glantz, 1980	79 (2)	39
Felson, 1982	74 (1)	34
MacArthur, 1982	114 (1)	28
Tyson, 1983	86 (4)	10
Avram, 1985	243 (2)	15
Thorn, 1985	120 (4)	<40
Murray, 1988	28 (1)	61
Morris, 1988	103 (1)	34
McGuigan, 1955	164 (1)	60
Welch, 1996	145 (1)	30

The table was modified from (Altman, 1998).

When asking why many scientific papers are wrong, even if statistics are correctly applied, we also have to consider that a study usually does not depend on a single experiment. We report on a number of experiments that, when taken together, make the case. Even if we achieve a significance level of 95% in each given experiment, when combined, the probability of an error increases steadily (Fig. 3).

The purpose of this article is not a review of statistics and statistical practice. It serves more as an illustration of yet another contributor to non-reproducibility of results. We might leave it with Andrew Lang: "He *uses statistics as a drunken man uses lamp-posts – for support rather than illumination*."

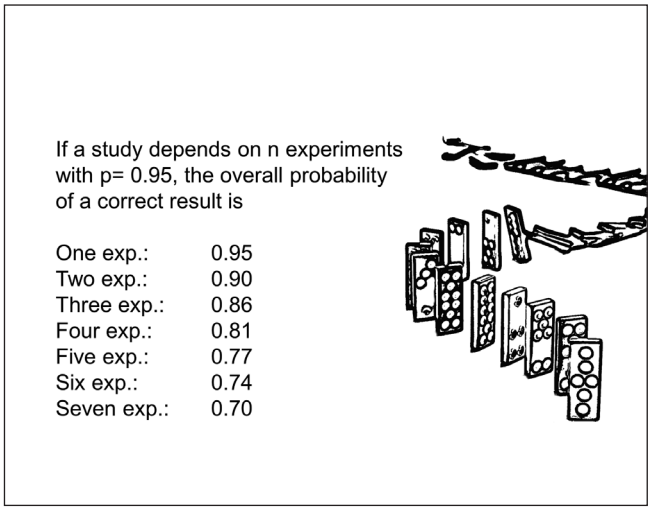

Fig. 3: If the overall conclusion of an article depends on a number of experiments, each with an error rate of 5%, the overall probability of a non-chance result decreases steadily

Consideration 7: Publication practices contribute to the misery – publish and perish, not publish or perish

The problem lies not only in the data generated, their statistical analysis, and the way we form an overall story from them: publication practices have their share in impeding objective science. In an interesting article, "*Why current publication practices may distort science*," Young et al. (2008) use an economic view on scientific publication behaviors: "*the small proportion of results chosen for publication are unrepresentative of scientists' repeated samplings of the real world. The self-correcting mechanism in science is retarded by the extreme imbalance between the abundance of supply (the output of basic science laboratories and clinical investigations) and the increasingly limited venues for publication (journals with sufficiently high impact). This system would be expected intrinsically to lead to the misallocation of resources. The scarcity of available outlets is artificial, based on the costs of printing in an electronic age and a belief that selectivity is equivalent to quality. Science is subject to great uncertainty: we cannot be confident now which efforts will ultimately yield worthwhile achievements. However, the current system abdicates to a small number of intermediates an authoritative prescience to anticipate a highly unpredictable future. In considering society's expectations and our own goals as scientists, we believe that there is a moral imperative to reconsider how scientific data are judged and disseminated*." The authors make a number of recommendations regarding how to improve the system:

"1. *Accept the current system as having evolved to be the optimal solution to complex and competing problems.*
2. *Promote rapid, digital publication of all articles that contain no flaws, irrespective of perceived "importance".*
3. *Adopt preferred publication of negative over positive results; require very demanding reproducibility criteria before publishing positive results.*
4. *Select articles for publication in highly visible venues based on the quality of study methods, their rigorous implementation, and astute interpretation, irrespective of results.*
5. *Adopt formal post-publication downward adjustment of claims of papers published in prestigious journals.*
6. *Modify current practice to elevate and incorporate more expansive data to accompany print articles or to be accessible in attractive formats associated with high-quality journals: combine the "magazine" and "archive" roles of journals.*
7. *Promote critical reviews, digests, and summaries of the large amounts of biomedical data now generated.*
8. *Offer disincentives to herding and incentives for truly independent, novel, or heuristic scientific work.*
9. *Recognise explicitly and respond to the branding role of journal publication in career development and funding decisions.*
10. *Modulate publication practices based on empirical research, which might address correlates of long- term successful outcomes (such as reproducibility, applicability, opening new avenues) of published* papers."

Please note that the authors' involvement with ALTEX, most recently with Peer Journal (https://peerj.com), and especially with the Evidence-based Toxicology Collaboration (http://www.ebtox.com) promotes some of these goals. While the former two foster digital open-access publication with new financial models reducing the costs to readers and authors, the series of Food for Thought … articles, commissioned t[4] white papers, and the systematic reviews under development in the EBTC aim to be exactly the "*critical reviews, digests, and summaries of the large amounts of biomedical data now generated*." The variety of initiatives for "*quality of study methods*" will add to this.

Consideration 8: The shortcomings of *in vitro* contributing to poor research

Earlier in this series of articles, the shortcomings of typical cell culture were discussed (Hartung, 2007). This article summed up experiences gained from the validation of *in vitro* systems and in the course of developing the Good Cell Culture Practice

guidance (Coecke et al., 2005). Six years later the arguments are largely the same: We do not manage to obtain *in-vivo*-like differentiation because we often use tumor cells (tens of thousands of mutations, loss and duplications of chromosomes), overpassage with selection of subpopulations, use non-physiologic culture conditions (hardly any cell contact, low cell density, no polarization, limited oxygen supply, non-homeostatic media exchange, temperature and electrolyte concentrations reflective of humans not rodents), force growth (fetal calf serum, growth factors), do not demand cell functions due to over-pampering, do not follow the *in vitro* kinetics giving consideration to the fate of test substances in the culture, and do not represent cell type interactions. For most aspects there are technical solutions, but few are applied, and if so, they are applied in isolation, solving some but not all of the problems. Beside this, there is a lack of quality control. If we take the estimates below, probably only 60% of studies use the intended cells without mycoplasma infection. Documentation practices in laboratories and publications are often lousy. There is some guidance (GLP increasingly adapted, GCCP see below) but it is rarely applied. The more recent mushrooming of cell culture protocol collections is an important step, but it is still not common to stick to them or at least to be clear in publications about deviations from them: We tend to toy around with the models until they work for us, and too often only for us.

There is some movement with regard to cell line authentication (see below). The earlier article summarizing the history and core ideas of GCCP (Hartung and Zurlo, 2012) did not address mycoplasma infection, a problem far from being solved. There are also some new aspects coming from the booming field of stem cells.

Hello, HeLa… – the cell you see more often than you would believe. Since 1967, cell line contaminations have been evident, i.e., another cell type was accidentally introduced into a culture and slowly took over. The most promiscuous so far are HeLa cells, actually the first human tumor cell line. The line was derived from cervical cancer cells taken on February 8, 1951, from Henrietta Lacks, a patient at Johns Hopkins. The cells have contributed to more than 60,000 research papers and the development of a polio vaccine in the 1950s (more on the interesting history in (Skloot, 2010)). Recently the HeLa genome has been sequenced (Landry et al., 2013) (please note some controversy around the paper which is currently being sorted out). It is most interesting to see the genetic make-up of the cells as summarized by Ewen Callawa in Nature[10]: "*HeLa cells contain one extra version of most chromosomes, with up to five copies of some. Many genes were duplicated even more extensively, with four, five or six copies sometimes present, instead of the usual two. Furthermore, large segments of chromosome 11 and several other chromosomes were reshuffled like a deck of cards, drastically altering the arrangement of the genes*." Do we really expect such a cell monster to show normal physiology? The cell line was found to be remarkably durable and prolific, as illustrated by its contamination of many other cell lines. It is assumed that, today, 10-20% of cell lines are actually HeLa cells and, in total, 18-36% of all cell lines are wrongly identified. Table 5 shows studies analyzing the problem over time extracted from (Hughes et al., 2007).

Tab. 5: Studies showing misidentified cell lines in various studies over time

1968:	100% (18) HeLa
1974:	45% (20) HeLa
1976:	30% (246) wrong (14% wrong species)
1977:	15% (279) wrong
1981:	about 100 contaminations in cells from 103 sources
1984:	35% (257) wrong
1999:	15% (189) wrong
2003:	15% (550) wrong
2007:	18% (100) wrong

Percentage of misidentified cell lines; total cell lines analyzed in brackets. These studies were extracted from (Hughes et al., 2007); c.f. references and more information.

A very useful list of such mistaken cell lines is available.[11] The problem has been raised several times (Macleod et al., 1999; Stacey, 2000; Buehring et al., 2004; Rojas et al., 2008; Dirks et al., 2010). A study (Buehring et al., 2004) from 2004 showed that HeLa contaminants were used unknowingly by 9% of survey respondents, likely underestimating the problem; only about a third of respondents were testing their lines for cell identity. More recently, a technical solution for cell line identification has been introduced by the leading cell banks (ATCC, CellBank Australia, sDSMZ, ECACC, JCRB, and RIKEN), i.e., short tandem repeat (STR) microsatellite sequences. STR are highly polymorphic in human populations, and their stability makes STR profiling (typing) ideal as a reference technique for identity control of human cell lines. We have to see how the scientific community takes this up. Isn't it a scandal that a large percentage of *in vitro* research is done on cells other than the supposed ones and misinterpreted this way?

Another type of contamination that is astonishingly frequent and has a serious impact on *in vitro* results is microbial infection, especially with mycoplasma (Langdon, 2003): Screening by the FDA for more than three decades showed that, of 20,000 cell cultures examined, more than 3000 (15%) were contaminated with mycoplasma (Rottem and Barile, 1993). Studies in Japan and Argentina reported mycoplasma contamination rates of 80% and 65%, respectively (Rottem and Barile, 1993). An analysis by the German Collection of Microorganisms and Cell Cultures (DSMZ) of 440 leukemia-lymphoma cell lines showed that 28% were mycoplasma positive (Drexler and Uphoff, 2002).

Laboratory personnel are the main sources of *M. orale*, *M. fermentans*, and *M. hominis*. These species of mycoplasma account for more than half of all mycoplasma infections in cell

[10] http://www.nature.com/news/most-popular-human-cell-in-science-gets-sequenced-1.12609

[11] http://www.hpacultures.org.uk/services/celllineidentityverification/misidentifiedcelllines.jsp

cultures and physiologically are found in the human oropharyngeal tract (Nikfarjam and Farzaneh, 2012). *M. arginini* and *A. laidlawii* are two other mycoplasmas contaminating cell cultures that originate from fetal bovine serum (FBS) or newborn bovine serum (NBS). Trypsin solutions provided by swine are a major source of *M. hyorhinis*. It is important to understand that the complete lack of a bacterial cell wall of mycoplasma implies resistance against penicillin (Bruchmüller et al., 2006), and they even pass 0.2 µm sterility filters, especially at higher pressure rates (Hay et al., 1989). Mycoplasma can have diverse negative effects on cell cultures (Tab. 6), and it is extremely difficult to eradicate this intracellular infection.

While there is good understanding in the respective fields of biotechnology, this is much less the case in basic research and mycoplasma testing is neither internationally harmonized with validated methods nor common practice in all laboratories on a regular basis. The recent production of reference materials (Dabrazhynetskaya et al., 2011) offers hope for the respective validation attempts. The problem lies in the fact that at least 20 different species are found in cell culture, though 5 of them appear to be responsible for 95% of the cases (Bruchmüller et al., 2006). For a comparison of the different mycoplasma detection platforms see (Lawrence et al., 2010; Young et al., 2010), and Table 7.

Tab. 6: Effects of mycoplasma contaminations on cell cultures

- Cell death and total culture degeneration and loss; increased sensitivity to apoptosis
- Alteration of cellular morphology
- Alteration of proliferation characteristics (growth, viability)
- Chromosomal aberrations (numerical and structural alterations); DNA fragmentation due to mycoplasma nucleases
- Alteration of cellular metabolism: Inhibition of cell metabolism; altered levels of protein, RNA and DNA synthesis with change of gene expression patterns
- Changes in cell membrane antigenicity (surface antigen and receptor expression)
- Interference with various biochemical and biological assays: Increase (or decrease) of virus propagation; reduction of transfection efficiencies; induction (or inhibition) of lymphocyte activation; induction (or suppression) of cytokine expression; influence on signal transduction; promotion of cellular transformation
- Specific effects on hybridomas: Inhibition of cell fusion; influence on selection of fusion products; interference in screening of monoclonal antibody reactivity; monoclonal antibody against mycoplasma instead of target antigen; reduced yield of monoclonal antibody; conservation of hybridoma

The table was combined from (Nikfarjam and Farzaneh, 2012) and (Drexler and Uphoff, 2002).

Tab. 7: Mycoplasma detection methods, their sensitivity, and advantages and disadvantages

Technique	Sensitivity	Pro	Con
Direct DNA stain	Low	Fast, cheap	Can be difficult to interpret
Indirect DNA stain with indicator cells	High	Easy to interpret because contamination amplified	Indirect and thus more time-consuming
Broth and agar culture	High	Sensitive	Slow (minimum 28d), can be difficult to interpret, problems of sample handling, lack of standards for calibration
PCR (endpoint and real-time-PCR)	High	Fast	Requires optimization, can miss low level infections, no distinction between live and dead mycoplasma
Nested PCR	High	Fast	More sensitive than direct PCR, but more likely to give false positives
ELISA	Moderate	Fast, reproducible	Limited range of species detected, reproducible
PCR ELISA	High	Fast, reproducible	May give false positives
Autoradiography	Moderate	Fast	Can be difficult to interpret
Immunostaining	Moderate	Fast	Can be difficult to interpret

This table was combined from (Garner et al., 2000, Young et al., 2010, Lawrence et al., 2010, Volokhov et al., 2011). Other less routinely used methods include microarrays, massive parallel sequencing, mycoplasma enzyme based methods, and recombinant cell lines.

Fig. 4: Karyograms of commercial CHO cells (parental) and CHO cells transfected with different CYP-450
Taken from the thesis work of Dr Alessia Bogni, co-supervised by Dr Sandra Coecke.

The advent of human embryonic and, soon after, induced pluripotent stem cells, appears to be something of a game changer. First it promises to overcome the problems of availability of human primary cells, though a variety of commercial providers nowadays make almost all relevant human cells available in reasonable quality but at costs that are challenging, at least for academia. We have to see, however, that we do not yet really have protocols to achieve full differentiation of any cell type from stem cells. This is probably a matter of time, but many of the non-physiologic conditions taken from traditional cell culture contribute here. Stem cells have been praised for their genetic stability, which appears to be better than for other cell lines, but we increasingly learn of their limitations in that respect too (Mitalipova et al., 2005; Lund et al., 2012; Steinemann et al., 2013). The limitations experienced first are costs of culture and slow growth; many protocols require months and labor, media, and supplement costs add up. The risk of infection unavoidably increases. Still we do not obtain pure cultures, often requiring a cell sorting, which, however, implies detachment of cells with the respective disruption of culture conditions and physiology.

Owing to the author's own experience with non-reproducible *in vitro* papers during his own PhD, in 1996 the author started an initiative toward Good Cell Culture Practice (GCCP), that led in 1999 to a workshop and declaration in the general assembly of the Third World Congress on Alternatives and Animal Use in the Life Sciences in Bologna, Italy. We then established an ECVAM working group and finally produced GCCP guidance (Coecke et al., 2005). The details of this process recently were summarized in this series of articles (Hartung and Zurlo, 2012). Here, only a single epiphany shall be added: in the PhD thesis of my student Alessia Bogni we obtained commercial CHO cell lines declared to be only transfected with single CYP-450 enzymes. The karyograms in Figure 4 show the dramatic effects with losses and fusions of chromosomes, some of which, in the lower right corner, could not even be identified. We would have interpreted any differences in experimental results only by the presence or absence of a single gene product…

GCCP acknowledges the inherent variation of *in vitro* test systems calling for standardization. GLP gives only limited guidance for *in vitro* (Cooper-Hannan et al., 1999) though some parts of GCCP have been adapted into a GLP advisory document by OECD for *in vitro* studies (OECD, 2004). The topic of quality of the publication of *in vitro* studies in journal articles also has been addressed in our Food for Thought … series earlier (Leist et al., 2010). GLP cannot normally be implemented in academia on the grounds of costs and lack of flexibility. For example, GLP requests that personnel be trained before they execute studies, while obviously students are "trained on the job." We hope that GCCP also will be guidance for journals and funding bodies, thereby enforcing the use of these quality measures.

GCCP guidance was developed before the broad use of human stem cells. We attempted an update in a workshop, which, strangely, never has been published but was made available as a manuscript on the ECVAM website[12]: "*hESC Technology for Toxicology and Drug Development: Summary of Current Status and Recommendations for Best Practice and Standardization. The Report and Recommendations of an ECVAM Workshop. Adler et al. Unpublished report*." We currently are aiming for an update workshop in early 2014 teaming up with FDA and the UK Stem Cell Bank.

[12] Available at: http://ihcp.jrc.ec.europa.eu/our_labs/eurl-ecvam/archive-publications/workshop-reports (last accessed 9 June 2013)

Conclusions

Science is increasingly becoming aware of the shortcomings of its approaches. John Ioannidis has stirred us up with papers like those entitled "*Why Most Published Research Findings Are False*" (Ioannidis, 2005b) ("*for many current scientific fields, claimed research findings may often be simply accurate measures of the prevailing bias*") or "*Contradicted and Initially Stronger Effects in Highly Cited Clinical Research*" (Ioannidis, 2005a). As early as 1994 Altman wrote on "*The scandal of poor medical research*" (Altman, 1994). This does not even address the contribution of fraud (Fang et al., 2012). These early warnings now have been substantiated with the unsuccessful attempts by industry to reproduce important basic research. Drummond Remmie phrased it like this: "*Despite this system, anyone who reads journals widely and critically is forced to realize that there are scarcely any bars to eventual publication. There seems to be no study too fragmented, no hypothesis too trivial, no literature too biased or too egotistical, no design too warped, no methodology too bungled, no presentation of results too inaccurate, too obscure, too contradictory, no analysis too self serving, no argument too trifling or too unjustified, and no grammar and syntax too offensive for a paper to end up in print. The function of peer review, then, may be to help decide not whether but where papers are published*."

The situation is not very different whether this is *in vitro* or *in vivo* work, which now often is combined anyway. Similar things can be said about *in silico* work (Hartung and Hoffmann, 2009), which is not only limited by the *in vitro* and *in vivo* data it is based on (trash in, trash out), but inherent problems of lack of data accuration and overfitting.[13] "*Torture numbers, and they'll confess to anything*" (Gregg Easterbrook). One difference is that *in vitro* approaches have developed the principles of validation. There is no field more self-critical than the area of alternative methods, where we spend half to one million $ and, on average, ten years to validate a method. Basic research could learn from this, not to go to the same extreme, which is becoming increasingly as much a burden as it is a solution to the problem, but to put sufficient effort into establishing the reproducibility and relevance of our methods. We are not calling for GLP for academia, but for the spirit of GLP to be embraced.

While this series of articles focuses mostly on toxicology, here we have attempted to extend some critical observations to research in general. This shall first of all show that toxicology is not different in its problems, and is perhaps even advanced with regard to internationally harmonized methods and quality assurance. It is perhaps too easy to just criticize. Henri Poincaré said "*To know how to criticize is good, to know how to create is better*." A simple piece of advice: The changes that clinical research has undergone should be adopted by basic research and regulatory sciences, especially weighing of evidence, documentation, and quality assurance. Publish less, but of better quality, or as Altman (1994) put it: "*We need less research, better research, and research done for the right reasons*."

References

Altman, D. G. (1994). The scandal of poor medical research. *BMJ 308*, 283-284.

Altman, D. G. (1998). Statistical reviewing for medical journals. *Stat Med 17*, 2661-2674.

Altman, D. G. (2002). Poor-quality medical research – what can journals do? *JAMA 287*, 2765-2767.

Andersen, B. (1990). *Methodological errors in medical research: An Incomplete Catalogue* (288pp). Chicago, USA: Blackwell Science Ltd.

Arrowsmith, J. (2011a). Trial watch: Phase II failures: 2008-2010. *Nat Rev Drug Discov 10*, 328-329.

Arrowsmith, J. (2011b). Trial watch: Phase III and submission failures: 2007-2010. *Nat Rev Drug Discov 10*, 87.

Arrowsmith, J. (2012). A decade of change. *Nat Rev Drug Discov 11*, 17-18.

Baker, D. D., Lidster, K. K., Sottomayor, A. A., and Amor, S. S. (2012). Reproducibility: Research-reporting standards fall short. *Nature 492*, 41.

Basketter, D. A., Clewell, H., Kimber, I., et al. (2012). A roadmap for the development of alternative (non-animal) methods for systemic toxicity testing – t^4 report. *ALTEX 29*, 3-91.

Bebarta, V., Luyten, D., and Heard, K. (2003). Emergency medicine animal research: does use of randomization and blinding affect the results? *Acad Emerg Med 10*, 684-687.

Begley, C. G. and Ellis, L. M. (2012). Drug development: Raise standards for preclinical cancer research. *Nature 483*, 531-533.

Bekelman, J. E., Li, Y., and Gross, C. P. (2003). Scope and impact of financial conflicts of interest in biomedical research a systematic review. *JAMA 289*, 454-465.

Blakey, D., Galloway, S. M., Kirkland, D. J., and MacGregor, J. T. (2008). Regulatory aspects of genotoxicity testing: from hazard identification to risk assessment. *Mutat Res 657*, 84-90.

Bottini, A. A. and Hartung, T. (2009). Food for thought ... on the economics of animal testing. *ALTEX 26*, 3-16.

Bottini, A. A. and Hartung, T. (2010). The economics of animal testing. *ALTEX 27, Spec Issue 1*, 67-77.

Briel, M., Müller, K. F., Meerpohl, J. J., et al. (2013). Publication bias in animal research: a systematic review protocol. *Syst Rev 2*, 23.

Bruchmüller, I., Pirkl, E., Herrmann, R., et al. (2006). Introduction of a validation concept for a PCR-based Mycoplasma detection assay. *Cytotherapy 8*, 62-69.

[13] A wonderful illustration by David J. Leinweber, Caltech: Stupid data miner tricks: overfitting the S&P 500, showing that the stock market behavior over 12 years could be almost perfectly explained by three variables, i.e., butter production in Bangladesh, United States cheese production, and sheep population in Bangladesh and United States; available at: http://nerdsonwallstreet.typepad.com/my_weblog/files/dataminejune_2000.pdf

Buehring, G. C., Eby, E. A., and Eby, M. J. (2004). Cell line cross-contamination: how aware are mammalian cell culturists of the problem and how to monitor it? *In Vitro Cell Dev Biol Anim 40*, 211-215.

Ciccone, A. and Candelise, L. (2003). Risk of cerebral haemorrhage after thrombolysis: Systematic review of animal stroke models. In *Proceedings of the European Stroke Conference*, Valencia, May 2003.

Coecke, S., Balls, M., Bowe, G., et al. (2005). Guidance on good cell culture practice. A report of the second ECVAM task force on good cell culture practice. *ATLA 33*, 261-287.

Cook, D. J., Mulrow, C. D., and Haynes, R. B. (1997). Systematic reviews: synthesis of best evidence for clinical decisions. *Ann Intern Med 126*, 376-380.

Cooper-Hannan, R., Harbell, J., and Coecke, S. (1999). The principles of Good Laboratory Practice: application to in vitro toxicology studies. *ATLA 27*, 539-577.

Corpet, D. E. and Pierre, F. (2003). Point: From animal models to prevention of colon cancer. Systematic review of chemoprevention in min mice and choice of the model system. *Canc Epidemiol Biomarkers Prev 12*, 391-400.

Corpet, D. E. and Pierre, F. (2005). How good are rodent models of carcinogenesis in predicting efficacy in humans? A systematic review and meta-analysis of colon chemoprevention in rats, mice and men. *Eur J Cancer 41*, 1911-1922.

Dabrazhynetskaya, A. A., Volokhov, D. V. D., David, S. W. S., et al. (2011). Preparation of reference strains for validation and comparison of mycoplasma testing methods. *J Appl Microbiol 111*, 904-914.

Dirks, W. G., Macleod, R. A. F., Nakamura, Y., et al. (2010). Cell line cross-contamination initiative: An interactive reference database of STR profiles covering common cancer cell lines. *Int J Cancer 126*, 303-304.

Drexler, H. G. and Uphoff, C. C. (2002). Mycoplasma contamination of cell cultures: Incidence, sources, effects, detection, elimination, prevention. *Cytotechnology 39*, 75-90.

Ersson, C., Moller, P., Forchhammer, L., et al. (2013). An ECVAG inter-laboratory validation study of the comet assay: inter-laboratory and intra-laboratory variations of DNA strand breaks and FPG-sensitive sites in human mononuclear cells. *Mutagenesis 28*, 279-286.

Fang, F. C., Steen, R. G., and Casadevall, A. (2012). Misconduct accounts for the majority of retracted scientific publications. *Proc Natl Acad Sci U S A 109*, 17028-17033.

Fellows, M. D., O'Donovan, M. R., Lorge, E., and Kirkland, D. (2008). Comparison of different methods for an accurate assessment of cytotoxicity in the in vitro micronucleus test. II: Practical aspects with toxic agents. *Mutat Res 655*, 4-21.

Fellows, M. D., Boyer, S., and O'Donovan, M. R. (2011). The incidence of positive results in the mouse lymphoma TK assay (MLA) in pharmaceutical screening and their prediction by MultiCase MC4PC. *Mutagenesis 26*, 529-532.

Festing, M. (2003). The need for better experimental design. *Trends Pharmacol Sci 27*, 341-345.

Fowler, P., Smith, K., Young, J., et al. (2012a). Reduction of misleading ("false") positive results in mammalian cell genotoxicity assays. I. Choice of cell type. *Mutat Res 742*, 11-25.

Fowler, P., Smith, R., and Smith, K. (2012b). Reduction of misleading ("false") positive results in mammalian cell genotoxicity assays. II. Importance of accurate toxicity measurement. *Mutat Res 747*, 104-117.

Garner, C. M., Hubbold, L. M., and Chakraborti, P. R. (2000). Mycoplasma detection in cell cultures: a comparison of four methods. *Br J Biomed Sci 57*, 295–301.

Goldacre, B. (2012). *Bad Pharma* (448pp). London, UK: Fourth Estate.

Gollapudi, B. B., Schisler, M. R., McDaniel, L. P., and Moore, M. M (2012). Reevaluation of the U.S. National Toxicology Program's (NTP) mouse lymphoma forward mutation assay (MLA) data using current standards reveals limitations of using the program's summary calls. *Toxicologist 126*, 448 (abstract).

Hackam, D. G. and Redelmeier, D. A. (2006). Translation of research evidence from animals to humans. *JAMA 296*, 1731-1732.

Hackam, D. G. (2007). Translating animal research into clinical benefit. *BMJ 334*, 163-164.

Harrell, F. E. (2011). Reproducible research. Presentation available at: https://www.ctsacentral.org/sites/default/files/documents/berdrrhandout.pdf (accessed on June 9, 2013)

Hartung, T. (2007). Food for thought ... on cell culture. *ALTEX 24*, 143-152.

Hartung, T. (2008a). Food for thought … on animal tests. *ALTEX 25*, 3-9.

Hartung, T. (2008b). Food for thought ... on alternative methods for cosmetics safety testing. *ALTEX 25*, 147-162.

Hartung, T. (2009a). Food for thought ... on evidence-based toxicology. *ALTEX 26*, 75-82.

Hartung, T. (2009b). Toxicology for the twenty-first century. *Nature 460*, 208-212.

Hartung, T. (2009c). Per aspirin ad astra... *ATLA 37, Suppl 2*, 45-47.

Hartung, T. and Hoffmann, S. (2009). Food for thought ... on in silico methods in toxicology. *ALTEX 26*, 155-166.

Hartung, T. (2010). Evidence-based toxicology – the toolbox of validation for the 21st century? *ALTEX 27*, 253-263.

Hartung, T. and Zurlo, J. (2012). Food for thought ... Alternative approaches for medical countermeasures to biological and chemical terrorism and warfare. *ALTEX 29*, 251-260.

Hay, R. J., Macy, M. L., and Chen, T. R. (1989). Mycoplasma infection of cultured cells. *Nature 339*, 487-488.

Helsel, D. R. (2006). Fabricating data: How substituting values for nondetects can ruin results, and what can be done about it. *Chemosphere 65*, 6.

Hoffmann, S. and Hartung, T. (2006). Toward an evidence-based toxicology. *Hum Exp Toxicol 25*, 497-513.

Hooijmans, C. R., Leenaars, M., and Ritskes-Hoitinga, M. (2010). A gold standard publication checklist to improve the quality of animal studies, to fully integrate the Three Rs, and to make systematic reviews more feasible. *ATLA 38*, 167-182.

Horn, J., de Haan, R. J., Vermeulen, M., et al. (2001). Nimodipine in animal model experiments of focal cerebral

ischemia: a systematic review. *Stroke 32*, 2433-2438.

Hughes, P., Marshall, D., Reid, Y., et al. (2007). The costs of using unauthenticated, over-passaged cell lines: how much more data do we need? *BioTechniques 43*, 575-584.

Ioannidis, J. P. A. (2005a). Contradicted and initially stronger effects in highly cited clinical research. *JAMA 294*, 218-226.

Ioannidis, J. P. A. (2005b). Why most published research findings are false. *PLoS Med 2*, e124-e124.

Ioannidis, J. P. A. (2012). Extrapolating from animals to humans. *Sci Transl Med 4*, 151.

Kilkenny, C. C., Browne, W. J., Cuthill, I. C., et al. (2010). Improving bioscience research reporting: the ARRIVE guidelines for reporting animal research. *PLoS Biol 8*, e1000412.

Kirkland, D., Aardema, M., Henderson, L., and Müller, L. (2005). Evaluation of the ability of a battery of three in vitro genotoxicity tests to discriminate rodent carcinogens and non-carcinogens. *Mutat Res 584*, 1-256.

Kirkland, D., Pfuhler, S., Tweats, D., et al. (2007). How to reduce false positive results when undertaking in vitro genotoxicity testing and thus avoid unnecessary follow-up animal tests: Report of an ECVAM Workshop. *Mutat Res 628*, 31-55.

Kirkland, D. (2010a). Evaluation of different cytotoxic and cytostatic measures for the in vitro micronucleus test (MNVit): Introduction to the collaborative trial. *Mutat Res 702*, 139-147.

Kirkland, D. (2010b). Evaluation of different cytotoxic and cytostatic measures for the in vitro micronucleus test (MNVit): Summary of results in the collaborative trial. *Mutat Res 702*, 135-138.

Knight, A. (2007). Systematic reviews of animal experiments demonstrate poor human clinical and toxicological utility. *ATLA 35*, 641-659.

Landry, J. J. M., Pyl, P. T., Rausch, T., et al. (2013). The genomic and transcriptomic landscape of a HeLa cell line. *G3 (Bethesda),* Epub March 11, 2013. doi: 10.1534/g3.113.005777.

Lang, T. (2004). Twenty statistical errors even you can find in biomedical research articles. *Croat Med J 45*, 361-370.

Langdon, S. P. (2003). Cell culture contamination: An overview. *Methods Mol Med 88*, 309-318.

Lawrence, B., Bashiri, H., and Dehghani, H. (2010). Cross comparison of rapid mycoplasma detection platforms. *Biologicals 38*, 6.

Lazarou, J., Pomeranz, B. H., and Corey, P. N. (1998). Incidence of adverse drug reactions in hospitalized patients. *JAMA 279*, 1200-1205.

Leist, M., Efremova, L., and Karreman, C. (2010). Food for thought ... considerations and guidelines for basic test method descriptions in toxicology. *ALTEX 27*, 309-317.

Leist, M. and Hartung, T. (2013). Reprint: Inflammatory findings on species extrapolations: humans are definitely no 70-kg mice. *ALTEX 30*, 227-230.

Lexchin, J. (2003). Pharmaceutical industry sponsorship and research outcome and quality: systematic review. *BMJ 326, 1167-1170*.

Lorge, E., Hayashi, M., Albertini, S., and Kirkland, D. (2008). Comparison of different methods for an accurate assessment of cytotoxicity in the in vitro micronucleus test. I. Theoretical aspects. *Mutat Res 655*, 1-3.

Lucas, C., Criens-Poublon, L. J., Cockrell, C. T., and de Haan, R. J. (2002). Wound healing in cell studies and animal model experiments by low level laser therapy; were clinical studies justified? A systematic review. *Lasers Med Sci 17*, 110-134.

Lumbreras, B., Parker, L. A., Porta, M., et al. (2009). Overinterpretation of clinical applicability in molecular diagnostic research. *Clin Chem 55*, 786-794.

Lund, R. J., Närvä, E., and Lahesmaa, R. (2012). Genetic and epigenetic stability of human pluripotent stem cells. *Nat Rev Genet 13*, 732-744.

Macleod, R. A. F., Dirks, W. G., Matsuo, M., et al. (1999). Widespread intraspecies cross-contamination of human tumor cell lines arising at source. *Int J Cancer 83*, 555-563.

Macleod, M. and van der Worp, H. B. (2010). Animal models of neurological disease: are there any babies in the bathwater? *Pract Neurol 10*, 312-314.

MacCallum, C. J. (2010). Reporting animal studies: good science and a duty of care. *PLoS Biol 8*, e1000413.

Mapstone, J., Roberts, I., and Evans, P. (2003). Fluid resuscitation strategies: a systematic review of animal trials. *J Trauma Injury Infect Crit Care 55*, 571-589.

Mignini, L. E. and Khan, K. S. (2006). Methodological quality of systematic reviews of animal studies: a survey of reviews of basic research. *BMC Med Res Methodol 6*, 10.

Mitalipova, M. M., Rao, R. R., Hoyer, D. M., et al. (2005). Preserving the genetic integrity of human embryonic stem cells. *Nat Biotechnol 23*, 19-20.

Moore, T. J., Cohen, M. R., and Furberg, C. D. (2007). Serious adverse drug events reported to the Food and Drug Administration, 1998-2005. *Arch Intern Med 167*, 1752-1759.

Moreno-Villanueva, M., Pfeiffer, R., Sindlinger, T., et al. (2009). A modified and automated version of the "Fluorimetric Detection of Alkaline DNA Unwinding" method to quantify formation and repair of DNA strand breaks. *BMC Biotechnology 9*, 39.

Moreno-Villanueva, M., Eltze, T., Dressler, D., et al. (2011). The automated FADU-assay, a potential high-throughput in vitro method for early screening of DNA breakage. *ALTEX 28*, 295-303.

NRC – National Research Council, Committee on Animal Models for Assessing Countermeasures to Bioterrorism Agents (2011). *Animal Models for Assessing Countermeasures to Bioterrorism Agents* (1-153). Washington, DC, USA: The National Academies Press. http://dels.nationalacademies.org/Report/Animal-Models-Assessing-Countermeasures/13233

Nikfarjam, L. and Farzaneh, P. (2012). Prevention and detection of Mycoplasma contamination in cell culture. *Cell J 13*, 203-212.

OECD (2004). Advisory document of the working group on GLP – the application of the principles of GLP to in vitro studies. *Series on Principles of Good Laboratory Practice*

and Compliance Monitoring 14, 1-18.

Olson, H., Betton, G., Robinson, D., et al. (2000). Concordance of the toxicity of pharmaceuticals in humans and in animals. *Regul Toxicol Pharmacol 32*, 56-67.

Petticrew, M. and Davey Smith, G. (2003). Monkey business: what do primate studies of social hierarchies, stress, and the development of CHD tell us about humans? *J Epidemiol Community Health 57, Suppl. 1*, A1-21.

Perel, P., Roberts, I., Sena, E., et al. (2007). Comparison of treatment effects between animal experiments and clinical trials: systematic review. *BMJ (Clinical research ed) 334*, 197.

Pfuhler, S., Kirkland, D., Kasper, P., et al. (2009). Reduction of use of animals in regulatory genotoxicity testing: Identification and implementation opportunities – Report from an ECVAM workshop. *Mutat Res 680*, 31-42.

Pfuhler, S., Kirst, A., Aardema, M., et al. (2010). A tiered approach to the use of alternatives to animal testing for the safety assessment of cosmetics: Genotoxicity. A COLIPA analysis. *Regul Toxicol Pharmacol 57*, 315-324.

Pottenger, L. H., Bus, J. S., and Gollapudi, B. B. (2007). Genetic toxicity assessment: employing the best science for human safety evaluation Part VI: When salt and sugar and vegetables are positive, how can genotoxicity data serve to inform risk assessment? *Toxicol Sci 98*, 327-331.

Pound, P., Ebrahim, S., Sandercock, P., et al. (2004). Where is the evidence that animal research benefits humans? *BMJ 328*, 514-517.

Prinz, F., Schlange, T., and Asadullah, K. (2011). Believe it or not: how much can we rely on published data on potential drug targets? *Nat Rev Drug Discov 10*, 712.

Roberts, I., Kwan, I., Evans, P., and Haig, S. (2002). Does animal experimentation inform human healthcare? Observations from a systematic review of international animal experiments on fluid resuscitation. *BMJ 324*, 474-476.

Rojas, A., Gonzalez, I., and Figueroa, H. (2008). Cell line cross-contamination in biomedical research: a call to prevent unawareness. *Acta Pharmacologica Sinica 29*, 877-880.

Rottem, S. and Barile, M. F. (1993). Beware of mycoplasmas. *Trends Biotechnol 11*, 143-151.

Sena, E. S., van der Worp, H. B., Bath, P. M. W., et al. (2010). Publication bias in reports of animal stroke studies leads to major overstatement of efficacy. *PLoS Biol 8*, e1000344.

Seok, J., Warren, H. S., Cuenca, A. G., et al. (2013). Genomic responses in mouse models poorly mimic human inflammatory diseases. *Proc Natl Acad Sci U S A 110*, 3507-3512.

Skloot, R. (2010). *The Immortal Life of Henrietta Lacks* (402pp). Crown Publishing Group; Reprint edition (February 2, 2010).

Skrabanek, P. and McCormick, J. (1990). *Follies and Fallacies in Medicine* (147 pp). Prometheus Books.

Snyder, R. D. and Green, J. W. (2001). A review of the genotoxicity of marketed pharmaceuticals. *Mutat Res 488*, 151-169.

Stacey, G. N. (2000). Cell contamination leads to inaccurate data: we must take action now. *Nature 203*, 356.

Steinemann, D., Göhring, G., and Schlegelberger, B. (2013). Genetic instability of modified stem cells – a first step towards malignant transformation? *Am J Stem Cells 2*, 39-51.

van der Worp, H. B., Howells, D. W., Sena, E. S., et al. (2010). Can animal models of disease reliably inform human studies? *PLoS Med 7*, e1000245.

van der Worp, H. B. and Macleod, M. R. (2011). Preclinical studies of human disease: time to take methodological quality seriously. *J Mol Cell Cardiol 51*, 449-450.

Volokhov, D. V., Graham, L. J., Brorson, K. A., and Chizhikov, V. E. (2011). Mycoplasma testing of cell substrates and biologics: Review of alternative non-microbiological techniques. *Mol Cell Probes 25*, 69–77.

Walmsley, R. M. (2008). GADD45a-GFP GreenScreen HC genotoxicity screening assay. *Expert Opin Drug Metab Toxicol 4*, 827-835.

Ward, D. J., Martino, O. I., Simpson, S., and Stevens, A. J. (2013). Decline in new drug launches: myth or reality? Retrospective observational study using 30 years of data from the UK. *BMJ Open 3*, e002088.

Young, N. S., Ioannidis, J. P. A., and Al-Ubaydli, O. (2008). Why current publication practices may distort science. *PLoS Med 5*, e201.

Young, L., Sung, J., Stacey, G., and Masters. J. R. (2010). Detection of Mycoplasma in cell cultures. *Nat Protoc 5*, 929-934.

Acknowledgement

Discussions with friends and colleagues shaped many of the arguments made here, especiallly Dr Marcel Leist, Dr Rosalie Elesprue, the GCCP taskforce, and the collaborators in the NIH transformative research grant "Mapping the Human Toxome by Systems Toxicology" (RO1ES020750) and FDA grant "DNTox-21c Identification of pathways of developmental neurotoxicity for high throughput testing by metabolomics" (U01FD004230) as well as NIH "A 3D model of human brain development for studying gene/environment interactions" (U18TR000547).

Correspondence to

Thomas Hartung, MD PhD
Center for Alternatives to Animal Testing
Johns Hopkins Bloomberg School of Public Health
615 North Wolfe Street
W7032, Baltimore, MD 21205, USA
e-mail: thartung@jhsph.edu