Vor- und Nachdenkliches ... zu Tierversuchen

Thomas Hartung
EC Joint Research Centre, IHCP/ECVAM, Ispra, Italy

Der vierte Artikel dieser Serie befasst sich mit dem Ziel, für das alle Alternativmethoden entwickelt werden, nämlich dem Tierversuch. Es ist nicht einfach, hier etwas Neues beizutragen, da Tierversuchsgegner im Laufe der Zeit viele Argumente zusammengetragen haben. Überraschenderweise gibt es relativ wenige relevante Übersichtsartikel in der wissenschaftlichen Literatur (MedLine). Ich hoffe, dass das nicht bedeutet, dass die Wissenschaftswelt nur wenig daran interessiert ist, die inhärenten Probleme dieses fundamentalen Werkzeugs der Lebenswissenschaften zu diskutieren. Tierversuche haben Vor- und Nachteile im Vergleich mit anderen Versuchsansätzen. In nur wenigen Fällen stellen sie den einzigen vernünftigen Ansatz dar. Zunehmend erlauben es moderne Methoden, Tierversuche zu reduzieren, zu verbessern und zu ersetzen (3R Prinzip), wie dies, wo immer möglich, von der europäischen Gesetzgebung gefordert wird. Dieser Artikel fasst typische Limitationen von Tierversuchen zusammen. Dennoch bedeutet das nicht, dass jede der einzelnen Limitationen auf alle Tierversuche zutrifft oder dass deren Alternativen weniger Limitationen aufweisen; dazu mehr in (Hartung, 2007b).

Eine ernsthafte Beurteilung der jeweils verwendeten Modelle ist notwendig, um Evidenz-basiert Schlussfolgerungen zu ziehen und Entscheidungen zu fällen. Nur die gründliche Beurteilung der Leistungsmerkmale eines jeden Werkzeugs, das in der Forschung eingesetzt wird, erlaubt es, Ergebnisse zu interpretieren und die eingesetzte Methode sinnvoll zu ergänzen. Die Öffentlichkeit hat zunehmend Bedenken, was das Wohlergehen von Tieren im Allgemeinen betrifft (z.B. Eurobarometer zu „Sozialen Werten, Wissenschaft und Technologie“ vom Juni 2005, in dem mehr als vier von fünf EU Bürgern angaben, dass die Gemeinschaft die Pflicht habe Tierrechte zu schützen, unabhängig von den damit verbundenen Kosten). Die EU bekannte sich auf internationaler Ebene zur Wichtigkeit des Schutzes von Tieren, die zu wissenschaftlichen Zwecken verwendet werden, und trat 1999 der Konvention des Europarats zum Schutz von Vertebraten bei, die für experimentelle und andere wissenschaftliche Zwecke verwendet werden, ETS 123 (Ratsentscheidung 1999/575/EC).

Hypothese 1: Ohne eine allgemeine Philosophie zur Tötung von Tieren, müssen wir mit fallweisen Entscheidungen, Uneindeutigkeit und Uneinigkeit leben

Auch in der Wissenschaft gibt es Fragen um Leben und Tod; Leben und Tod von Patienten und von Tieren im Tierversuch zum Beispiel. Die Grenze zwischen dem, was erlaubt oder zumindest fraglich ist, Leben zu gefährden oder Leben zu nehmen, hängt von Zweck und Umständen ab. Tabelle 1 versucht dieses Konzept einzufangen; es zeigt, dass wir die Tötung derselben Lebewesen als Schädlinge oder Gesundheitsrisiken gutheißen, die wir in der Forschung schützen wollen (mit geringfügigen geographischen Unterschieden). Solcher Schutz tendiert dazu, zu expandieren (zum Beispiel in der laufenden Revision des europäischen Gesetzes zum Schutz von Labortieren, in der embryonale Stadien im letzten Drittel der Entwicklung und einige Weich- und Schalentiere eingeschlossen werden sollen). Diese Entwicklung kann als Parallelent-

Tab. 1: Die Lizenz zum Töten ist zweckabhängig

	Insekten	Fische/ Vögel	Nager	Andere Säuger	Nichtmenschliche Primaten	Menschen-affen	Menschen
Pflanzenschutzmittel/ Biozide							
Nahrungsmittel & Produkte							
Forschung				EU US			verboten
Selbstverteidigung							

Weiss = generell akzeptiert; grau = entfällt; schwarz = beschränkt in der EU und den USA. Die Abbildung versucht darzustellen, dass die Tötung von Tieren und Menschen für verschiedene Zwecke allgemein akzeptiert wird – starke Beschränkungen gibt es nur im Gebiet der Forschung, mit bemerkenswerten Unterschieden besonders zwischen der Gesetzgebung der EU und der USA.

wicklung zur Ausbreitung humanitärer Prinzipen im Allgemeinen angesehen werden.

Es ist offenkundig, dass Töten ein Teil der Natur ist. Es ist offenkundig, dass wir buchstäblich mit jedem unserer Schritte kleine Tiere töten könnten. Es ist offenkundig, dass wir Tiere konsumieren, dass wir uns und unsere Nahrung schützen, dass wir Tiere opfern, um neue Heilmittel für Erkrankungen zu finden. Es gibt also keine Option, die „kein Töten" lautet, daher müssen wir eine Grenze definieren. In einem Kontinuum evolutionärer Komplexität (von Amöben bis zum Menschen) müssen Schwellen definiert werden, welche Spezies geschützt werden sollen und in welchem Umfang. Der sinnvollste Ansatz scheint zu sein, uns auf das zu beschränken, was unausweichlich und rechtfertigbar ist – was natürlich jede Entscheidung utilitaristisch und von den Umständen abhängig macht. So können wir gesellschaftliche und technische Entwicklungen vorantreiben, welche Tierversuche vermeiden und weniger rechtfertigbar machen. Das ist im Einklang mit dem Sinn des 3R Ansatzes, einer andauernden Aufforderung, weiter zu ersetzen, zu verbessern und zu reduzieren. Eine weiterführende ethische Debatte würde den Rahmen dieses Artikels sprengen. Jedoch bildet die Fähigkeit von Tieren, zu leiden, eine weitgehend akzeptierte Grundlage (Jeremy Bentham, *An Introduction to the Principles of Morals & Legislation*, 1789) für die Anerkennung von Tieren als empfindungsfähige Wesen und einen soliden Ausgangspunkt. Die europäische Wissenschaftsstiftung hat es kürzlich folgendermassen formuliert: „Labortiere haben nicht nur instrumentalen Wert, sondern auch einen intrinsischen Wert, der respektiert werden muss. Tiere müssen immer als empfindungsfähige Lebewesen behandelt werden." (*European Science Foundation, Policy Briefing*, September 2000, eigene Übersetzung).

Hypothese 2: Kein Modell ist perfekt

Der Ausdruck „Modell" deutet bereits auf eine Abweichung von der Realität hin; meist ist dies eine Vereinfachung, eine Reduktion von Variablen. Das einzige perfekte Experiment in den Biowissenschaften, das etwas über die Reaktion von Menschen aussagen würde, wäre an einer großen und diversen Gruppe von Menschen unter natürlichen Expositionsbedingungen – für eine prospektive Studie selten eine realistische Option. Die Abweichungen von diesem theoretischen Idealfall sind unausweichlich, wenn es weiter Fortschritt geben soll, aber wir vergessen oft, welche Kompromisse wir eingegangen sind, um einen praktikablen Ansatz zu ermöglichen. Die Limitationen von *in vitro* Modellen wurden kürzlich zusammengefasst (Hartung, 2007b). Auch Tiermodelle haben Limitationen, wenn sie Wirkungen auf die menschliche Gesundheit widerspiegeln sollen, z.B. in der Toxikologie, aus der hier Schlüsselbeispiele angeführt sind. Die Lage ist bei der Arzneimittelentwicklung nicht anders. Der bekannte NIH Pharmakologe Bernhard Brodie formulierte es schon 1964 wie folgt: „…es ist meist reines Glück, dass Tierversuche zu klinisch nützlichen Arzneimitteln führen."

Unterschiede zwischen Spezies – Nur selten kann dasselbe Experiment sowohl an Menschen als auch an Tieren durchgeführt werden; wo dies möglich ist (z.B. Hautirritation (York et al., 1996; Basketter et al., 2004)) wurde eine Korrelation von ca. 60% festgestellt. Einfacher ist es, verschiedene Labortierspezies dem gleichen Experiment zu unterziehen. Sogar zwischen verschiedenen Nagerspezies werden typischerweise Korrelationen von nur 70% festgestellt, einige Zitate sind bei Hoffmann and Hartung (2006) zu finden, wobei bekanntlich die meisten Labortierspezies einander viel mehr ähneln als dem Menschen. Es gibt keinen Grund anzunehmen, dass irgendeine Spezies Wirkungen auf die menschliche Gesundheit besser widerspiegeln würde als Wirkungen auf irgendeine andere Spezies. Schon die Bioverfügbarkeit, d.h. wie viel der oral gegebenen Ausgangsmenge aufgenommen wird und eine Wirkung haben kann, unterscheidet sich wesentlich zwischen verschiedenen Spezies, wie überzeugend von Grass und Sinko gezeigt (2002, Abb. 1, Seite 437). Sogar Affen können beträchtlich in ihrer Reaktion von der des Menschen abweichen, wie wir tragischerweise 2006 feststellen mussten, als der TeGenero anti-CD28 Antikörper, nachdem er bei einer 500-fach höheren Konzentration im Affen keine Nebenwirkungen aufwies, innerhalb weniger Stunden in sechs Freiwilligen zu Organversagen führte (Bhogal and Combes, 2006).

Pharmako- und Toxikokinetiken von Verbindungen unterscheiden sich zwischen Tier und Mensch – Menschen sind keine 70kg-Ratten: Es gibt gewaltige Unterschiede zwischen dem Menschen und allen Labortieren, z.B. was den metabolischen Umsatz, Nahrungs- und Wasseraufnahme, pH im Magen/Darm-Trakt, Körpertemperatur, Metabolismus von Xenobiotika bezüglich Kapazität, Geschwindigkeit und gebildeten Metaboliten, Verteilung von Körperfett, Eliminationswege, Proteinbindung / Carrierproteine, Abwehrmechanismen, Lebensdauer, usw. betrifft. Die Durchsicht von z.B. Michael Derelanko's „*Toxicologist's Pocket Handbook*" kann Augen öffnen. In der regulativen Toxikologie wird typischerweise ein Korrekturfaktor von 100 verwendet, d.h. es wird angenommen, dass Menschen bis zu 100-fach sensitiver sind als Labortiere, um Interspeziesunterschiede und interindividuelle Unterschiede zu berücksichtigen. Es wurde jedoch kürzlich gezeigt, dass sogar ein Faktor von 100 nicht immer ausreicht, und dass besonders sensitive Individuen hier nicht mitberücksichtigt sind (Wood, 1998, Falk-Filipsson et al., 2007; Bokkers and Slob, 2007). Es scheint eher angebracht von Unsicherheitsfaktoren zu sprechen, denn das sind sie tatsächlich (IGHRC, 2003). Jedoch berücksichtigt auch dieser Ansatz nur den quantitativen Fehler der Beurteilung, denn es ist nachvollziehbar, dass die Gefährdung das Wichtige ist und dass die quantitative Einschätzung letztlich komplexeren Erwägungen folgt. Nur wenig ist jedoch hinsichtlich des qualitativen Fehlers bekannt, d.h. die Zuordnung einer toxischen Wirkung oder nicht. Im Allgemeinen haben die interindividuellen Unterschiede in der Pharmakologie bis-

lang viel mehr Aufmerksamkeit genossen (Übersichtsartikel sind zum Beispiel Inaba et al., 1995; Lin, 2007) als in der Toxikologie (Lovell, 1993; Lipscomb, 2002), obwohl es keinen Grund gibt anzunehmen, dass sie hier eine geringere Rolle spielen.

Inzuchtstämme – Der breite Einsatz von Inzuchtstämmen hat geholfen, die Variabilität von experimentellen Ergebnissen zu verringern. Jedoch impliziert die Arbeit mit eineiigen Zwillingen auch, dass die Variabilität der (Tier-) Population nicht wiedergegeben wird (s. Festing, 1993; Lovell, 1993; Kacew and Festing, 1996).

Junge Tiere – Kosten und logistische Überlegungen begründen die Nutzung von jungen Tieren für Experimente; jedoch werden die meisten Gesundheitsbeeinträchtigungen in älteren Menschen beobachtet, welche auch an weiteren Grunderkrankungen leiden können. Kürzlich wurde sogar vorgeschlagen, noch weiter zu gehen und den Einfluss des Epigenoms, zu dem das Chromatin und die kovalente Modifikation von DNS durch Methylierung gehören, als Nahtstelle zwischen der dynamischen Umgebung und dem ererbten statischen Genom zu berücksichtigen (Szyf, 2007).

Beschränkung auf ein Geschlecht – Dies soll nicht als Aufruf zur standardmäßigen Testung in beiden Geschlechtern verstanden werden, aber wieder wird das Spektrum der Sensitivitäten verringert, und es sollte zumindest etwas Überlegung hinter der Wahl des Geschlechts stecken. Der Beitrag von geschlechtsspezifischen Unterschieden ist besser für die Wirkung von Medikamenten etabliert (zusammengefasst in Gandhi et al., 2004) als für toxische Wirkungen, aber es gibt keinen Grund anzunehmen, dass sie hier eine geringere Rolle spielen.

Gruppengröße – Meist wird die kleinstmögliche Gruppengröße für Tierversuche gewählt, oft ohne eine Poweranalyse durchzuführen, welche anzeigen würde, ob ein signifikantes Ergebnis überhaupt zu erwarten ist. Das Problem wird dadurch verstärkt, dass oft maximale Information durch die Erfassung vieler Endpunkte oder Zeitpunkte abgefragt wird, ohne die Statistik an die multiple Testung anzupassen; z.B. sehen Standard-Testprotokolle für Toxizität bei wiederholter Gabe die Erfassung von bis zu 40 Endpunkten vor. Die einfachste, aber konservative, Bonferroni-Korrektur zum Beispiel würde erfordern, das Signifikanzniveau vom typischen p=0,05 auf 0,00125 (d.h. 0,05 geteilt durch 40) für jeden Endpunkt zu reduzieren. Angenommen, dass eine toxische Wirkung bei wiederholter Gabe mit der typischen Gruppengröße von 10 (OECD Testprotokoll 407) bei p=0,05 gerade signifikant ist, dann bräuchten wir in diesem Beispiel 22 Tiere pro Gruppe, damit dieser Unterschied bei p=0,00125 im ungepaarten zweiseitigen t-Test signifikant würde. Es gibt gute Gründe, warum klinische Studien mit tausenden eingeschriebenen Patienten nur einen primären Endpunkt haben.

Kosten – Die Pflege einer ausreichenden Anzahl von Tieren und die Bezahlung von qualifizierten Mitarbeitern sind sehr kostenintensiv. Diese Kosten können sich auf mehrere zehn- bis hunderttausend Euro belaufen, z.B. bei höheren Primaten. Als Folge ist die Anzahl an Replikaten und Wiederholungen meist limitiert. Aber auch gewöhnliche Tierversuche können ziemlich teuer sein: Eine Umfrage zu den Testkosten in Europa (Fleischer, 2007) zeigte durchschnittliche Kosten von z.B. 1.200 € für Hautirritation, 1.350 € für Augenirritation, 50.000 € für die 28-Tage-wiederholte orale Gabe (doppelt soviel bei Inhalation), 330.000 € für eine Zwei-Generationen- Reproduktionstoxizitätsstudie oder 780.000 € für einen Karzinogenitäts-Bioassay in Ratten.

Unrealistische Dosierung und Expositionsdauer – Behandlungen mit hohen Dosierungen (in der Toxikologie meist die maximal tolerierte Dosierung, d.h. bis zu 10% der Tiere sterben) werden durchgeführt, obwohl das kaum die Exposition eines Menschen widerspiegeln würde (Sumner and Stevens, 1994; Mehendale, 1995). Obwohl es nicht viele Hinweise gibt, dass zusätzliche Effekte bei niedrigeren Dosierungen auftreten (zusammengefasst von Kamrin, 2007), treten viele zusätzliche aber irrelevante Wirkungen auf, wenn Gewebskonzentrationen extrem hoch sind und Abwehrmechanismen überwunden werden.

Viele Gesundheitheitsschädigungen sind multifaktoriell – aber die Tiermodelle, die verwendet werden, haben meist nur eine einzige (artifizielle) Krankheitsursache. Dies ist möglicherweise eingängiger bei Modellen menschlicher Erkrankungen, bei denen fast immer genetische, psychologische, soziale, Ernährungs- und Umweltrisikofaktoren eine wesentliche Rolle spielen. Bei der menschlichen Intoxikation spielen ebenfalls viele Kofaktoren eine Rolle, einschließlich Gesundheitsstatus, genetische Vorbelastung und Verhalten bei Exposition.

Tiere sind gestresst – Es gibt viele Ursachen für Stress von Tieren: Transport, zu wenig Zeit zur Anpassung, Käfige ohne „*enrichment*“ (z.B. Möglichkeiten, Spezies-spezifische Verhaltensweisen auszuführen) (Wolfer et al., 2004), Handhabung der Tiere, insbesondere Maßnahmen, die tagsüber bei nachtaktiven Spezies durchgeführt werden, Veränderungen in der Gruppe bei der Randomisierung, zu viele männliche Tiere in einem Käfig, um nur einige hervortretende Beispiele zu benennen.

Adaptive Anpassungen von Tieren werden unterschätzt – Dies ist vor allem für die beliebte Knock-out-Maus-Experimente relevant; wir vergessen oft, dass diese Tiere schon ihr ganzes Leben lang Zeit hatten, ihren Gendefekt zu kompensieren, aber wir vereinfachen unsere Interpretation als ob allein ein einziger Gendefekt vorliegen würde. Auch kann jede Exposition gegenüber Xenobiotika oder Traumata Abwehrmechanismen, Entzündung und Wundheilungsprozesse aktivieren, welche die Reaktion auf nachfolgende (toxische) Behandlungen beeinflussen.

Vertrauliche Daten – Vor allem im regulatorischen Kontext ist die Tatsache,

dass Daten zu Tierversuchen nicht zugänglich gemacht werden, eine enorme wissenschaftliche Limitation; dies geschieht häufig, um kommerzielle Interessen zu schützen (Informationen an die Konkurrenz weitergeben, Beschränkungen, um die Daten für die eigene Entwicklung zu verwenden), doch manchmal ist auch nur der Aufwand ein mangelnder Anreiz zur Publikation.

Akute Modelle von chronischen Phänomenen – Der Druck, Ergebnisse schnell zu liefern, ist enorm groß, was dazu führt, dass wir Tiermodelle von chronischen Erkrankungen entwickeln, die schnell Symptome erzeugen.

Post-hoc Definition der experimentellen Hypothese – vor allem in der Grundlagenforschung ist es üblich, die Hypothese den Ergebnissen anzupassen, um in der Publikation eine „Geschichte erzählen" zu können; wenige sind sich bewusst, dass dies Auswirkungen auf die zu verwendenden statistischen Verfahren hat (*a posteriori* Testung).

Tests sind entweder sensitiv oder spezifisch – diese zwei Eigenschaften schließen einander gegenseitig aus; so können wir uns entweder auf positive oder negative Ergebnisse eines Tests verlassen, doch akzeptieren wir meist beide Ergebnisarten von Tierversuchen mit dem gleichen Vertrauen (Hoffmann and Hartung, 2005).

Die Prävalenz toxikologischer Wirkungen wird bei der Interpretation von Ergebnissen aus Tierversuchen außer Acht gelassen – Sogar ein fast perfekter Tierversuch würde uns alleine nicht weiterhelfen. Der Grund dafür wird in der diagnostischen Medizin „Prävalenz" genannt. In einfachen Worten: Wenn ein Ereignis unwahrscheinlich ist, ist die Frequenz falsch-positiver Ergebnisse eines Tests ausschlaggebender für seine Verlässlichkeit – wenn ich nur wenig echt-positive Ergebnisse unter den positiven Ergebnissen habe, kann ich mich nicht auf die positiven Ergebnisse des Tests verlassen. Meist suchen wir seltene Eigenschaften von Verbindungen, sowohl bei der Wirkstoffsuche (eine Verbindung aus hundert oder tausend) als auch in der Toxikologie (2-10% der Verbindungen). Der Anteil von entweder toxischen oder effektiven Verbindungen in einer Gruppe von Verbindungen liegt daher meist im Bereich von wenigen Prozent, und auch Tests mit einer guten Spezifität werden einige Prozent falsch-positive Ergebnisse liefern, d.h. ein Anteil der gleich groß oder größer ist als der der echt-positiven Resultate. Wir haben bereits in einem früheren Artikel auf diese Problematik hingewiesen (Hoffmann and Hartung, 2005). Die typische Kombination von mehreren Tests, z.B. in verschiedenen Labortierspezies, durch die die Sensitivität erhöht werden soll, z.B. um keine toxischen Verbindungen zu übersehen, verringert die Spezifität und steigert dadurch weiter die Anzahl falsch-positiver Verbindungen.

Hypothese 3: Letalität in der akuten Toxizitätsprüfung könnte häufig nur eine Nebenwirkung sein

Die Mortalität von Tieren ist häufig das Ergebnis eines Mangels an Nahrungs- oder Wasseraufnahme, nicht der primären Wirkung einer Verbindung. Wenn kleine Nager keine Nahrung aufnehmen können, sterben sie innerhalb von Stunden – wahrscheinlich wären viele Verbindungen nicht toxisch, wenn eine einfache Zuckerlösung injiziert werden würde. Auf ähnliche Weise trägt der Verlust an Körperwärme wahrscheinlich zur Letalität vieler Behandlungen bei. Die Körpertemperatur wird durch einen Mangel an Bewegung (die zu mehr als 80% der Wärmeproduktion führt), ungünstige Oberflächen/Körpergewicht-Verhältnis von kleinen Tieren und Hypothermie statt einer Fieberreaktion bei Nagern beeinflusst. Wahrscheinlich ist häufig die Translokation von Bakterien aus dem Darm das, was die Tiere tatsächlich umbringt, vor allem bei oral verabreichten Verbindungen, welche die gastrointestinalen Barrieren beschädigen. Hier würde ich gerne eine Beispielrechnung vorstellen (basierend auf Ernst Rietschel, Borstel, persönliche Mitteilung), die auf den ersten Blick nicht viel mit Toxikologie zu tun zu haben scheint: Der Darm eines einzigen Menschen enthält ein Kilogramm Gram-negative Bakterien mit etwa 50 g Endotoxin. Diese Menge an Endotoxin wäre ausreichend, um durch Injektion eine Million Menschen zu töten oder in einer Milliarde Menschen Fieber zu erzeugen. Ratten und Mäuse sind 100-fach weniger sensibel gegenüber Endotoxin. Mir ist nicht viel über die Masse und Zusammensetzung von Nagerkot bekannt, aber sicherlich enthält er genug Endotoxin, um viele Nager „endogen" umzubringen. Dieses Beispiel berücksichtigt nicht einmal die Translokation von lebenden Bakterien mit anschließender Infektion, einer Hauptursache für Mortalität bei ähnlich traumatisierten und toxifizierten Patienten (Übersichtsartikel sind zum Beispiel Deitch et al., 1996; Stechmiller et al., 1997; Gatt et al., 2007).

Ich würde dieses Argument gerne umkehren, um, vor allem bei akuten oralen Intoxikationen, die Hypothese, dass die Tiere nicht an einer toxischen Wirkung sondern an sekundären Wirkungen im Darm versterben, zu belegen: Es wurde durch mindestens drei Großansätze dokumentiert, dass Zytotoxizität recht gut mit akuter oraler Toxizität korreliert (siehe Halle-Register, MEIC-Studie und kürzlich die ICCVAM/NICEATM/ECVAM Validierungsstudie). Tatsächlich macht es wenig Sinn, wenn wir annehmen, dass die Verbindungen mit komplexen Kinetiken adsorbiert, verteilt und metabolisiert werden und auf mehr als 400 verschiedene Gewebe mit unterschiedlichen Sensitivitäten wirken können. Könnte es sein, dass der Tierversuch lediglich die Zytotoxizität am Darmepithel misst, welche zur Translokation von Bakterien führt? Ironischerweise würde das bedeuten, dass wir diesen Tierversuch mit *in vitro* Ansätzen sehr gut vorhersagen könnten, weil der Tierversuch ein Phänomen abbildet (Zytotoxizität im Darm), das für Menschen irrelevant ist (wir würden erbrechen – was Nager nicht können – oder die Intoxikation entfernen bevor sie den Darm erreicht, Intensivmedizin

bereitstellen, usw.). Statt unserer 9 Millionen € Anstrengung A-Cute-Tox (http://www.acutetox.org/), könnte eine gut geplante Serie von Tierversuchen (Blasphemie vom Leiter von ECVAM…) beweisen, dass die Referenzmethode irrelevant ist. Bemühungen des Zentrums NC3R in England, die Forderung nach akuten Toxizitätsstudien für Arzneimittel abzustellen (Chapman and Robinson, 2007) – weil sie für Arzneimittel sinnlos sind – verdient unsere volle Unterstützung und sollte bald auf Chemikalien, Pflanzenschutzmittel und Kosmetika ausgeweitet werden.

Hypothese 4: Können Tierversuche überhaupt chronische, systemische Toxizität beim Menschen vorhersagen?

Natürlich bekommen wir aus Tierversuchen Vorhersagen, aber sind sie gut genug? Eine Vorhersage ist akzeptabel wenn ihre Unsicherheit akzeptabel ist. Doch um die Unsicherheit einschätzen zu können, müssten wir wissen, was die Wahrheit ist. Wie soll das möglich sein mit den riesigen Unterschieden zwischen menschlicher Intoxikation und der experimentellen Situation, welche oben beschrieben ist? Daneben ist die hauptsächliche Limitation für die Extrapolation zum Menschen typischerweise der Expositionsteil: Menschen sind gleichzeitig tausenden Verbindungen ausgesetzt und nur wenigen davon auf kontrollierte/kontrollierbare und kontinuierliche Weise – wie sollen wir einzelne Verbindungen, die ein Problem darstellen, erfassen? Im Falle von Tabak und dem Fokus auf eine beschränkte Anzahl von gesundheitlichen Auswirkungen wurde deutlich, wie schwer es ist, die Ursache einer toxischen Wirkung nachzuweisen. Wie soll dies mit weniger evidenten, schlechter nachweisbaren Ursachen erreicht werden? Wir werden immer nur das finden, was wir schon kennen („wykiwyg = *what you know is what you get*“), d.h. im besten Fall können wir einen positiven Befund in einem Tierversuch nachvollziehen. Bei der hohen Anzahl von falsch-positiven Ergebnissen von vorbeugenden Prüfungen und dem enormen Aufwand, den epidemiologische Studien kosten, ist es unwahrscheinlich, dass viele kausale Zusammenhänge gezeigt werden können. Es besteht kein Risiko für ihn, wenn ein Risikoprüfer eine Verbindung als gesundheitsgefährdend für den Menschen deklariert – wenn es keine zusätzlichen, unterstützenden Daten gibt, wird niemand den Aufwand betreiben, dies nachzuprüfen. Bei unklaren Expositionsszenarien wird niemand in der Lage sein, dies nachzuprüfen.

Es gibt einen Weg, die Verlässlichkeit von toxikologischen Prüfungen einzuschätzen: Der Vergleich von verschiedenen Labortierspezies. Dies wird einem nicht die Wahrheit offenbaren (Gesundheitsgefährdung für den Menschen), aber es wird einem eine gute Einschätzung der Verlässlichkeit der Ergebnisse im Allgemeinen liefern. Dieser Ansatz ist von Anfang an verzerrt, da verschiedene Nagerspezies, welche typischerweise benutzt werden, näher mit einander verwandt sind als mit dem Menschen, aber er ist immerhin richtungweisend. Mit der bemerkenswerten Ausnahme des Karzinogenese-Bioassays, gibt es wenige derartige Vergleiche. Dies betrifft besonders die Toxizität bei wiederholter Gabe. Die Datenbank, die vom deutschen Fraunhofer Institut in Hannover (Bitsch et al., 2006) eingerichtet wurde, stellt eine Ausnahme und eine wertvolle Ressource für derartige Vergleiche dar.

Wenn wir das Konzept der Unsicherheitsfaktoren weiter ausbauen (Tab. 2), können wir beginnen, über die Größenordnung der Unsicherheit nachzudenken: Die Zahlen mögen angezweifelt werden (oder, noch besser, mit besseren Schätzungen ersetzt werden), aber eine grobe Berechnung würde vorschlagen, dass 50% unserer (qualitativen) Gefährdungseinschätzungen falsch sind und dass wir auf der quantitativen Seite um mehr als das 1.250-fache hinsichtlich der Potenz eines Toxins daneben liegen können. Stellt das eine akzeptable Unsicherheit dar?

Hypothese 5: Wirtschaftliche Faktoren beeinflussen die Verwendung von Tieren und den Ersatz von Tierversuchen

Tierversuche durchzuführen ist ein Geschäft: Es gibt zu vielen Schlüsselaspekten in diesem Sektor keine offiziellen, zusammengefassten Daten, aber die Studie von Prognos (http://ec.europa.eu/environment/chemicals/lab_animals/ia_en.htm), die im Auftrag der Europäischen Kommission anlässlich der Revision der Direktive für den Schutz von Labortieren durchgeführt wurde, schätzte, dass ein Umsatz von etwa 3 Milliarden € pro Jahr in der EU mit Tierversuchen erzielt wird; ca. 1.330 Einrichtungen (Industrie, Auftragsforschung und Universitäten) in der EU führen Tierversuche durch. Die Zahl der Züchter und Anbieter von Tieren für experimentelle Zwecke liegt aber wahr-

Tab. 2: Die toxikologische Unsicherheitskette

	Reproduzierbarkeit des Tierversuchs	Spezies Extrapolation	Inter-individuelle Unterschiede	Resultierende Gefahr für den Menschen	Unsicherheit der Exposition	Resultierendes Risiko für den Menschen
Quantitativ = Dosis-Wirkung (Unsicherheitsfaktor)	**1,25** C.V. typischerweise 20-30%	**10**	**>10**	**>125**	**>10** (persönliche Schätzung)	**>1250**
Qualitativ = Gefahr oder nicht (% richtig identifiziert)	**90%**	**60-70%** sens. 90% spez. 10-20%	**90%** (persönliche Schätzung mehr quantitativ)	**53%**		

scheinlich im Bereich von nur ein paar Dutzend. Ungefähr 9.300 neue Projekte, die Tierversuche beinhalten, starten jedes Jahr in der EU. Es wurde geschätzt, dass ein durchschnittliches Projekt ungefähr 300.000 € über drei Jahre kostet. Das alleine ist ein riesiger Markt mit Geschäftsinteressen, Lobbyisten, usw. Es ist eine Branche mit den natürlichen Mechanismen des Widerstands auf den Druck zur Veränderung, aber dies betrifft auch die Menschen, die in diesem Gebiet arbeiten: Ihre Fertigkeiten, ihre Grundlage für persönliche Bedeutung und Ruf, stehen unter dem Druck zur Veränderung. Jemand, der ein Testlabor für Tierversuche unterhält, wird versuchen einen Weg zu finden, dieses weiter betreiben zu können. Ein Experte der Planung, Durchführung und Interpretation von Tierversuchen wird eher dazu tendieren, diese Werkzeuge weiterhin anzuwenden. Im Bereich der Sicherheit ist es besonders schwierig, Einspruch zu erheben: Es ist nicht leicht, die Sicherheitsbedenken des Toxikologen in Frage zu stellen – wer würde für Zweifel an Bedenken zur Sicherheit die Verantwortung übernehmen wollen? Das Dezimieren von Verbindungen (durch vorbeugende Testplanung und Testinterpretation) verstärkt diese psychologische Machtbasis nur.

Gleichzeitig liegen Sicherheitstests auch im Interesse von Firmen: Zunächst einmal haben sie ein starkes Interesse, sichere Produkte auf den Markt zu bringen, ferner hilft dies aber auch, formal ihre Haftbarkeit zu verringern („alles Menschen mögliche wurde getan, um die Sicherheit zu gewährleisten“ – sarkastisch formuliert, die Segnung nach der rituellen Opferung von Tieren, oder die moderne Variante des mittelalterlichen Ablassbriefes). Ein dritter Faktor könnte es sein, dass die erhöhten Entwicklungskosten, die mit Sicherheitsprüfungen einhergehen, große Firmen vor kleinen bevorteilen; ein ähnlicher Prozess ereignete sich in der Arzneimittelentwicklung, wobei hier die klinischen Studien die Kostensituation so weit trieben, dass es sich nun nur noch ein Dutzend Firmen leisten können, neue Produkte bis zur Markteinführung zu entwickeln. Mit den viel kleineren Gewinnmargen der chemischen Industrie könnten ähnliche Auswirkungen durch die Forderung von Tierversuchen für Chemikalien, welche sich bei High-Production-Volume (HPV)-Chemikalien auf mehrere Millionen € belaufen können, erzielt werden.

Einen weiteren wirtschaftlichen Faktor stellen die nicht-Zoll-Handelshemmnisse dar, die durch unterschiedliche Erfordernisse an Sicherheitsprüfungen in verschiedenen wirtschaftlichen Gebieten erzeugt werden. Zum Glück haben die Europäische Vereinigung, OECD und WTO dieses Problem angegangen; trotzdem werden die globalen Märkte zur anhaltenden Verwendungen von traditionellen Methoden führen, bis auch der letzte wichtige Markt auf Alternativen umgestiegen ist (Bottini et al., 2007).

Hypothese 6: Es ist Glaube und nicht Wissenschaft, der unsere Erwartungen bezüglich der Validität vieler Tierversuche bestimmt

Seien wir provokativ: Ein Gebiet, dass auf Glauben beruht, ähnelt eher der Religion als einer Wissenschaft. Ein Gebiet, das mit dem Stand der Dinge zufrieden ist, riskiert es, ein Glaubenssystem zu werden. Wie Richard Dawkins es formulierte: „Eine der wirklich schlimmen Auswirkungen der Religion ist, dass sie uns lehrt, dass es eine Tugend ist, zufrieden zu sein, etwas nicht zu verstehen“ (Dawkins 2006, S.126, eigene Übersetzung). Das Vertrauen auf traditionelle Ansätze und der Widerstand gegen die Anwendung von modernen mechanistischen und technologischen Ansätzen sind Eigenschaften eines religiösen Ansatzes. Es ist ver-lockend (durch den Ersatz von „Religion“), das Zitat von Oscar Wilde anzupassen „Wahrheit, im Falle der (Toxikologie), ist nur die Meinung, die überlebt hat“.

Oberflächliche Validität *versus* Validierung – Wir tendieren dazu, an die Relevanz von Tierversuchen zu glauben, wenn einige ihrer Eigenschaften unseren Erwartungen entsprechen; im Gegensatz dazu bedeutet Validierung, die Relevanz eines Modells systematisch in Frage zu stellen. Ein interessantes Beispiel könnte das Magengeschwür sein (Lee, 2000). So lange wir daran glaubten, dass die Überproduktion von Säuren der zentrale Pathomechanismus war, verwendeten wir Tierversuche, die dies widerspiegelten und das erwartete Magengeschwür hervorriefen. Als gezeigt wurde, dass *Helicobacter pylori* die Hauptursache von Magengeschwüren ist, wurden neue Modelle etabliert. Das bedeutet, dass die oberflächliche Validität, d.h. die offensichtliche Widerspiegelung unseres Verständnisses der Pathogenese, die treibende Kraft hinter der Wahl und dem Glauben an das entsprechende Tiermodell war.

Standardisierung *versus* Tradition – Obwohl Standardisierung eine Voraussetzung für die gegenseitige Anerkennung von Daten ist, vor allem im regulatorischen Kontext (Bottini et al., 2007), und auch notwendig ist für die Ansammlung von genügend Wissen über das Modell, muss sie von einem „so haben wir es immer gemacht“ Ansatz (Tradition) abgegrenzt werden. Dieses Konzept deutet auf das Dilemma hin, dass die dauerhafte Festlegung einer Methode dazu führt, dass Innovation verpasst wird, wobei ein Mangel an Standardisierung die gegenseitige Anerkennung verhindert. Das bedeutet, dass Modelle angepasst werden müssen, wenn es dafür einen triftigen Grund gibt. Dies muss allerdings gut dokumentiert und gleichzeitig allgemein verbreitet werden, um einen neuen Standard zu etablieren. So wie die Standardisierung eine Voraussetzung für die Validierung ist, ist es das Vertrauen, das durch Validierung entsteht, welches erst die Basis für die Veränderung von Standardmethoden bildet.

Die meisten Testprotokolle für Tierversuche sind nicht validiert – Die systematische Beurteilung von Leistungsmerkmalen (Reproduzierbarkeit, Detektionsgrenze, Relevanz der Ergebnisse, mechanistische Grundlage) von Tierversuchen fehlt typischerweise. Obwohl dies nun größtenteils für die Einführung neuer Tierversuche, welche mittlerweile validiert werden müssen, anerkannt wird (Bravo!), wird dies für alle akzeptierten, traditionellen Tests

nicht in Frage gestellt. Offenbar gibt es große Befürchtungen, dieses Thema anzuschneiden und frühere Bewertungen in Frage zu stellen. Wir könnten unseren „Stand der Unschuld“ verlieren, aufgrund dessen wir behaupten können, dass alle Bewertungen mit bestem Wissen und Gewissen durchgeführt wurden.

Die richtige Zielspezies ist kein Garant für Relevanz – In der Ökologie, der Ökotoxikologie oder der Veterinärmedizin wird häufig angenommen, dass die Testergebnisse relevant sind, weil sie in der Zielspezies durchgeführt werden. Eine weitere „intellektuelle Abkürzung“ oder ein bequemer Glaube: Mit etwa 400.000 Fischspezies gibt es wahrscheinlich kein minder großes Problem, die Sensitivität und Spezifität jeder Spezies zu modellieren, als es das für Säugetiere gibt. Die meisten Limitationen, die bereits angesprochen wurden, treffen auch zu, wenn in der Zielspezies getestet wird.

Lebewesen sind unglaublich komplex – aber wir nutzen sie, um einfache Fragen zu beantworten. Jede Interpretation ist eine übermäßige Vereinfachung von dem, was tatsächlich alles im Experiment abgelaufen ist und unsere Kapazität zur Interpretation übersteigt. So folgen wir dem Prinzip „wykiwyg“ („*what you know is what you get*“, siehe oben, aber vielleicht sollten wir hier sagen, „*what you believe is what you get*“), d.h. Ergebnisse von Experimenten aufgrund unseres momentanen Verständnisses der Pathophysiologie zu interpretieren.

Ausblick

Wir leben in einer unvollkommenen Welt und haben unvollkommene Werkzeuge. Wie sind die Aussichten, trotzdem relevante Wissenschaft zu machen, den Verbraucher zu schützen und neue Arzneimittel mit einer hohen Wahrscheinlichkeit zu identifizieren? Zuerst müssen wir die Limitationen unserer Werkzeuge verstehen und diese wo möglich optimieren, standardisieren und validieren (Hartung, 2007a), was uns zu einem Evidenz-basierten Ansatz führen kann (Hoffmann and Hartung, 2006). Das bekannte Zitat von Bertrand Russel, wie er Gott erkläre, warum er nicht glaube, „Nicht genügend Beweise, Gott, nicht genügend Beweise“, könnte auf einige Gebiete der in Frage zu stellenden toxikologischen Tradition angepasst werden. Zweitens dürfen wir uns nicht allein auf ein Werkzeug verlassen: Die Kombination von Werkzeugen mit komplementären Stärken, wie einem sensitiven und einem spezifischen Test, kann Schwachstellen bewältigen. Integriertes Testen, vor allem unter Nutzung verschiedener Technologien, ist besonders vielversprechend. Dennoch werden die Werkzeuge, wie Teststrategien systematisch zusammengesetzt werden und mit denen diese validiert werden können, erst entwickelt. Drittens, müssen wir aufhören zu glauben und anfangen zu zweifeln. Es gibt eine Tendenz, *a priori* an die Relevanz von Tiermodellen zu glauben („*in vivo veritas*“). Diese wird durch einen Drang, eine definitive Aussage zu treffen, weitergetrieben (aufgrund von wirtschaftlichem Druck und Druck zur Publikation). Glauben ist wahrscheinlich der unwissenschaftlichste Ansatz.

Der Ausblick? 3R! Doch sind wahrscheinlich die einfachsten Ansätze von den aktuellen *in vitro* und verbesserten *in vivo* Techniken größtenteils schon entwickelt worden. Was *in silico* Technologien, welche der Validierung immer näher kommen, beisteuern können, muss noch ergründet werden (einige Gedanken zu diesem Thema in einem der nächsten „Vor- und Nachdenkliches…“ Artikel). Doch könnte eine wirklich neue Qualität durch die Kombination von Ansätzen unter Nutzung von (*high-throughput* und *high-content*) *in vitro* Modellen mit Bioinformatik und mechanistischen Informationen in Ansätzen der Systembiologie erreicht werden. Dieser Pfad wurde vor kurzem durch ein wichtiges visionäres Dokument durch die US National Academy of Sciences (*Committee on Toxicity and Assessment of Environmental Agents*, 2007) entwickelt. Aber Vision ist auch wieder ein eher religiöser Ansatz…

Danksagung

Die fortdauernden Diskussionen mit Kollegen und Freunden, vor allem den Mitarbeitern von ECVAM, die diese Ideen geprägt haben, werden dankbar anerkannt. Limitationen von Tiermodellen wurden kürzlich auch für die Proceedings vom 2007 LIMPE Seminar „*Experimental Models in Parkinson’s disease*“, Alghero, Sardinia, Italy, September 23 - 25, 2007 (Hartung T. Thoughts on limitations of animal models. Parkinsonism & Related Disorders 2008, in press) zusammengefasst.

Literatur

Literaturverzeichnis siehe „References“ Seite 8f.

Korrespondenzadresse

Prof. Dr. Dr. Thomas Hartung
EU Joint Research Centre
Institute for Health and Consumer Protection
European Centre for the Validation of Alternative Methods (ECVAM)
21020 Ispra
Italien
E-Mail: thomas.hartung@ec.europa.eu