Thomas Hartung, ECVAM

Vor- und Nachdenkliches... zur Validierung

Vor- und Nachdenken?

In der Internet-Ära eine wissenschaftliche Zeitschrift herauszugeben ist eine grosse Herausforderung: Häufig werden Artikel über Datenbankrecherchen abgerufen, mit der Folge, dass immer weniger Wissenschaftler die Zeitschrift tatsächlich in den Händen halten und ihren Inhalt durchsehen. Viele verpassen dabei, was eine wissenschaftliche Zeitschrift von einer Artikelsammlung unterscheidet: das Austauschforum einer bestimmten Fachgruppe. Mit meiner neuen Reihe "*Food for thought*" (Vor- und Nachdenken) möchte *ALTEX* diese Form der Kommunikati0on stärken, indem eine sehr persönliche Sichtweise zu Themen im Bereich der Ersatzmethoden zum Tierversuch vorgestellt wird.

Als erstes Thema habe ich die Kernaufgabe von ECVAM herausgesucht: die Validierung. Ich möchte einige Gedanken dazu vorstellen, offene Probleme und Lösungsansätze besprechen, aber keine endgültigen Antworten liefern. Kommentare und Anregungen sind sehr willkommen, wie auch aktive Beiträge, um Lösungen für diese offenen Probleme zu entwickeln.

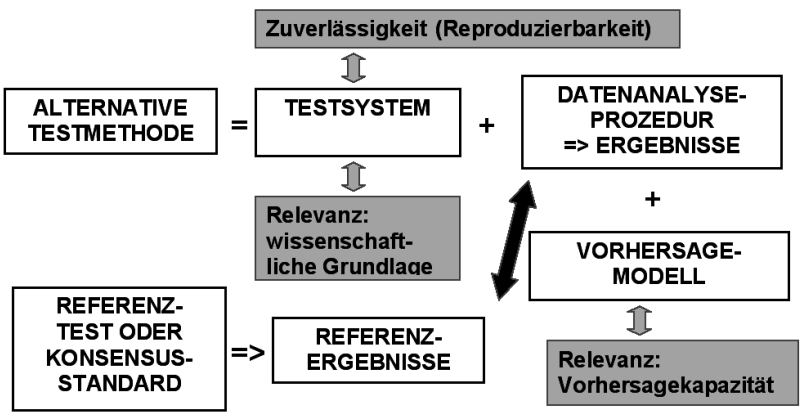
Was bedeutet die Validierung von Alternativmethoden?

Laut Wörterbuch wird der Begriff „alternativ" erst seit den Siebziger Jahren mit der Bedeutung „besser als das Bestehende" verwendet. Das ist tatsächlich eine gute Umschreibung des Kerns der Wissenschaft selbst: Fortschritt zu Besserem. Daher ist es eigentlich bemerkenswert, dass dieser Begriff heute allgemein so eng mit Ersatzmethoden zu Tierversuchen verbunden ist. In diesem Gebiet streben wir eine humanere Wissenschaft an, eine Wissenschaft, die das Leiden von Tieren vermindert oder verhindert. Bemerkenswert ist, dass dieses Gebiet auch Vorreiter des Schutzes der wissenschaftlichen Qualität ist. Validierung von Alternativmethoden bedeutet zunächst, die Relevanz der wissenschaftlichen Methoden nachzuweisen. Das ist ein erstaunlich junges Konzept, das sich erst in den letzten 20 Jahren entwickelt hat. Tatsächlich entstand es in genau diesem Gebiet der Biomedizin, im Bereich der Alternativmethoden.

Es gibt zwei Hauptgründe, weshalb gerade dieser Bereich eine Vorreiterfunktion übernommen hat: Zum Einen steht im Zusammenhang mit (regulatorischen) Sicherheitsprüfungen sehr viel auf dem Spiel (unsere Gesundheit, mögliche Skandale) und zum Andern bewegen wir uns in dem wissenschaftlichen Gebiet mit den wahrscheinlich meisten „traditionellen" (um nicht zu sagen „altmodischen") Ansätzen. Wo sonst in der Wissenschaft nutzen wir heute noch Methoden, die sich in den letzten 60 Jahren kaum geändert haben?

Diese zwei Aspekte bedingen einander: Wo die Sicherheit von Konsumenten oder Patienten betroffen ist, ist es schwierig, neue Ansätze durchzusetzen; die Regulatoren haben guten Grund, konservativ zu handeln. Um den Herausforderungen der Selbsterneuerung einerseits gerecht zu werden und gleichzeitig Sicherheitsstandards zu bewahren wurde die formale Validierung eingeführt. Dadurch soll die Tauglichkeit von neuen Methoden geprüft werden, und Veränderungen der Prüfpraxis werden nur zugelassen, wenn sie auf einer soliden Grundlage beruhen. In diesem Sinne sichert die Validierung das „Überleben des Besseren", das "*Survival of the Fittest*". So wird in einem Gebiet, in dem einige „Fossilien" geschützt werden, manchmal eben eine „nicht-natürliche Selektion" notwendig.

Das Prinzip der Validierung ist in Abbildung 1 dargestellt: Die Alternativmethode besteht aus dem Testsystem und einer Methode zur Datenanalyse (*data*



Validierung **ist ein Prozess, in dem die wissenschaftliche Grundlage und die Reproduzierbarkeit eines Testsystems und die Vorhersagekapazität eines zugehörigen Vorhersagemodells einer unabhängigen Evaluierung unterzogen werden.**

Abb. 1: Definition der Validierung

analysis procedure – DAP). Letztere ist ein Algorithmus, der verwendet wird, um aus den Rohdaten das Ergebnis des Tests zu berechnen. Der Referenztest dient dazu zum Vergleich. Wenn es keinen Referenztest gibt oder er nicht für den Vergleich angewendet werden kann, wird ein so genannter Konsensus-Standard erstellt, d.h. Experten einigen sich auf die Referenz, zum Beispiel eine Auswahl von positiven und negativen Chemikalien.

Die Ergebnisse des Alternativtests werden dann mit den Referenzergebnissen verglichen. Zu diesem Zweck wird meist ein Vorhersage-Modell benötigt, welches die Ergebnisse der Alternativmethode in die Kategorien oder Maßeinheiten der Referenzmethode umrechnet (z.B. wird ein Maß für Zytotoxizität in Toxizitätsklassen nach EU-Klassifizierung umgerechnet). Es gibt drei Kriterien für die Validität, die in der Abbildung durch graue Kästen unterlegt sind: (i) Reproduzierbarkeit des Testsystems, (ii) dessen wissenschaftliche, mechanistische Grundlage und (iii) dessen Vorhersagekapazität für die Referenzergebnisse. Ferner sind Qualitätskontrollen des Tests (Standardisierung, besonders auch die Definition seiner Anwendbarkeit) und der Validierungsprozess an sich (vor allem dessen Transparenz und Unabhängigkeit) Voraussetzungen und bilden zusammen den Validierungsprozess.

Einige häufige Missverständnisse bezüglich der Validierung

Missverständnis 1: Validierung ist eine Tierschutzinitiative

Die Validierung von Alternativmethoden ist in erster Linie ein Prozess der Qualitätskontrolle. Ihr Ziel ist es, die Anwendung von unausgereiften oder ungeeigneten Methoden in sensiblen Bereichen wie Sicherheitsprüfung und Produktentwicklung zu verhindern. Sie hat eine Türsteherfunktion inne und hat *per se* nicht zur Aufgabe, Alternativmethoden zu fördern. Tatsächlich ist nur eine winzige Anzahl aller wissenschaftlichen Methoden geeignet, gut etablierte Tierversuche zu ersetzen, und wieder nur ein Bruchteil derer sind ausreichend gut standardisiert, um in den Validierungsprozess starten zu können. Von diesen Methoden werden ungefähr ein Drittel in der Prävalidierung und ungefähr ein Drittel in der Validierungsphase durchfallen. Validierung bedeutet also die rigorose Aussortierung von vielen an sich wertvollen in vitro und seit neuerem *in silico* Ansätzen (Worth et al., 2004) aus der wissenschaftlichen Forschung, um die wenigen zu finden, die dann als validierte Ersatzmethoden betrachtet werden können.

Warum fördert Validierung dennoch Alternativen und stellt nicht nur ein Hindernis für ihre Anwendung dar? Die optimistische Antwort lautet: Weil Alternativmethoden einfach besser sind und sich gegen die konventionellen Tierversuche zur Toxizitätsprüfung oft behaupten können. Zitate wie „humane Wissenschaft ist die beste Wissenschaft" oder „Tierschutz und gute Wissenschaft sind nur zwei Seiten der gleichen Medaille" drücken dies aus.

Es gibt tatsächlich Grundlagen, die diese Ansicht stützen (Goldberg und Hartung 2006):

- Tierversuche widerspiegeln den wissenschaftlichen Ansatz der Zeit, in der sie entwickelt wurden, nicht unbedingt die heutige wissenschaftliche Herangehensweise und das heutige Verständnis.
- Tierversuche wurden nie einer gründlichen Qualitätskontrolle/Validierung unterzogen, basieren stattdessen eher auf Konvention; meist gibt es keine präzise Anleitung zur Durchführung der Methode.
- Tierversuche sind „under-powered", d.h. es werden vom Standpunkt eines Statistikers (Entschuldigung an die Tierliebhaber unter den Statistikern – das betrifft nur die professionelle Ansicht) viel zu wenig Tiere pro Experiment verwendet, um sichere Schlüsse ziehen zu können. Kosten, Arbeitszeit und Tierschutz beschränken den Gebrauch von Tieren in Hinsicht auf Gruppengrösse und Anzahl der Wiederholungen.

Letzterer Nachteil wird zu häufig durch die unsachgemäße Verwendung von Statistik verdeckt (der ewige Favorit ist natürlich der einseitige t-Test) oder gleich durch die Unterlassung statistischer Analysen. Wir nehmen gerne neue Endpunkte in Tierversuchen auf, ohne für die multiple Testung zu kompensieren: Eine Signifikanzgrenze von 5% bedeutet dann aber, dass einer aus 20 Endpunkten für eine negative Substanz falsch-positiv sein kann. Regulatorische SOPs schliessen häufig 40 Endpunkte ein. Da wird es schwierig, überhaupt noch eine negative Substanz zu finden…

Ein weiterer beliebter Trick, mit dem aus kleinen Gruppen Ergebnisse herausgepresst werden, ist, Inzuchtstämme in den Tierversuchen zu verwenden: Die Testung eineiiger Zwillinge eliminiert die Variabilität des Tests – prima, wenn der Endpunkt der Untersuchung nicht durch eben diese Variabilität beeinflusst wird. Aber wer kontrolliert das schon richtig?

Missverständnis 2: Validierung ist die Kalibrierung einer Methode

Die Bezeichnung „Validierung" wird in vielerlei Zusammenhängen verwendet. Bei chemischen und physikalischen Methoden oder generell in Qualitätskontrollprozessen von ISO/GLP bezieht sie sich hauptsächlich auf die Reproduzierbarkeit und die Definition der erforderlichen Kontrollen. Wissenschaftler, die in analytischen Bereichen arbeiten, verstehen häufig nicht den Umfang an Arbeit und Zeit, der benötigt wird, um eine Validierungsstudie für eine Toxizitätsprüfung durchzuführen. Die Kalibrierung von Methoden stellt sicher, dass wir Parameter richtig messen. Der Kern der Validierung von Alternativmethoden ist jedoch, festzustellen, ob wir die richtigen Parameter messen (d.h. seine Relevanz), und festzustellen, für welche Testmaterialien der Test verwendet werden kann (negativ ausgedrückt, die Limitationen des Tests zu erfassen). Es wäre wohl besser gewesen, einen ganz neuen Begriff für die Validierung von Alternativmethoden zu erfinden, zum Beispiel „Relevantisierung". Der Gebrauch des mehrdeutigen Begriffs „Validierung" führt dazu, dass zu viele Menschen zu früh denken, sie hätten den Prozess verstanden.

Missverständnis 3: Tierexperimente sind durch ihre erfolgreiche, langjährige Anwendung validiert worden

Die beste Antwort auf diese Aussage, die mir bisher untergekommen ist (obwohl sie ursprünglich in einem anderen Zusammenhang gemeint war), ist Folgende:

„Aus Erfahrung lernen ist vielleicht nichts anderes als lernen, die gleichen Fehler mit immer mehr Selbstvertrauen zu begehen.“ (Petr Skrabanek und James McCormick, *Follies and Fallacies in Medicine*, Tarragon Press, Glasgow, 1989).

Der ideale Weg, Tierversuche zu evaluieren, wäre deren Ergebnisse gegen epidemiologische Studien in Menschen zu stellen. Allerdings sind solche Studien meist nicht verfügbar oder können zumindest nicht mit vernünftigem Aufwand durchgeführt werden. Warum sollte Erfahrung, die nicht systematisch erhoben wurde, hier beitragen können? Es hat 50 Jahre gedauert, bis nachgewiesen werden konnte, dass Rauchen zu Krebs führt. Wie sollen wir effektiver mit allen anderen potentiell gefährlichen Substanzen umgehen, bei denen es deutlich schwieriger ist, die Exposition abzuschätzen? Die 10 bis 15 Jahre Latenzzeit zwischen der Exposition und der Diagnose von Krebs macht epidemiologische Studien extrem schwierig. Andere chronische gesundheitliche Beeinträchtigungen sind nicht einfacher zu beurteilen – vielleicht vergeht weniger Zeit zwischen der Exposition und dem Auftreten von chronischer, systemischer Toxizität oder Fortpflanzungstoxizität, aber die möglichen Manifestationen sind auch viel diverser (Prieto et al., 2006).

Wir könnten fragen, ob chronische Toxizität überhaupt vorhergesagt werden kann. Eine Welt mit ungefähr 140.000 synthetischen Chemikalien, die uns ermöglichen, fast 100 Jahre alt zu werden, kann für unsere Gesundheit nicht so schädlich sein. Diesen Umstand der erfolgreichen Eliminierung bestimmter Chemikalien gutzuschreiben, scheint anmaßend, da eine systematische Risikoabschätzung neuer Chemikalien (in Europa seit 1981) lediglich mit den letzten 4.700 durchgeführt wurde.

Die Behauptung, dass Tierversuche uns vor den gefährlichen Wirkungen der Chemikalien beschützen, ist schwer zu entkräften: Die meisten Substanzen sind sowieso ungefährlich, andere sind noch nicht lang genug auf dem Markt, um Probleme zu verursachen, und in den meisten Fällen kann kein kausaler Zusammenhang zwischen Exposition und Wirkung wissenschaftlich bewiesen werden. Es ist frustrierend, dass nur akute und topische toxische Wirkungen einen Vergleich zwischen den Ergebnissen von Tierversuchen und menschlichen Daten einfach erlauben, und natürlich sind es auch gerade diese, die relativ leicht durch Alternativmethoden ersetzt werden können.

Missverständnis 4: Prävalidierung ist was vor der Validierung abläuft

Die Prävalidierung wurde als wichtiger Kontrollpunkt für den Einstieg in grosse Ringstudien in das Validierungsschema aufgenommen (Hartung und Spielmann, 1995). Allerdings sind die Standardisierung von Methoden und die Evaluierung ihrer Reproduzierbarkeit integrale Bestandteile des Validierungsprozesses. Tatsächlich werden diese in anderen Bereichen oft schon als die Validierung verstanden (siehe Missverständnis 2). Die Arbeit, die in dieser Phase durchgeführt wird, dauert häufig sogar länger und ist aufwändiger als die eigentliche, endgültige Validierungsphase, wird aber dennoch oft nur als Vorbereitungsphase verstanden. Wir haben daher in der Definition des modularen Ansatzes (Hartung et al., 2004) weitgehend auf diesen Ausdruck verzichtet. Er wird aber sicherlich weiterhin einen integralen Schritt in der Organisation einer prospektiven Validierungsstudie darstellen.

Missverständnis 5: Omik-Technologien werden uns schnell neue Testmöglichkeiten eröffnen, es wird aber schwer sein, diese zu validierten

Es gibt grosse Hoffnungen für die Entwicklung von neuen Alternativmethoden aus dem Bereich der Toxikogenomik, -proteomik und -metobonomik. Wie so häufig ruft Technologie, die schwer zu verstehen ist, oft unbegründete Hoffnungen oder Ängste hervor. Tatsächlich ist es beim Thema Omik-Technologien als Alternativen genau anders herum: Validierung von Omik-Technologien kann schnell durchgeführt werden, aber diese Technologien sind schwer zu standardisieren, und es bleibt zu klären, ob sie überhaupt angemessene Mittel darstellen, um die Antworten, die wir brauchen, zu liefern. Dennoch laufen Bemühungen, die Anwendung dieser Technologien im regulatorischen Rahmen zu ermöglichen (Corvi et al., 2006). Die Hauptprobleme sind folgende:

- Die Messung von vielen (datenreichen) Endpunkten verbessert nicht die Qualität des Testsystems, mit dem man beginnt („Wo Müll reingeht, kommt auch Müll raus”).
- Die hochkomplexen Prozesse sind schwer zu standardisieren und Qualitätskontrolle ist schwierig.
- Die Technologien sind noch immer zu aufwändig, um sie in Routine-Laboratorien einsetzen zu können.

Es gibt wenig Zweifel daran, wie sie validiert werden müssen, wenn sie ausreichend weit entwickelt sind (Corvi et al., 2006): sie werden mit einigen Substanzen herausgefordert, und die Reproduzierbarkeit und Vorhersagekapazität werden geprüft.

Dennoch versprechen diese Techniken, effektiver neue Biomarker zu identifizieren, mit denen die gesuchte Wirkung erfasst werden und ein tieferes mechanistisches Verständnis gewonnen werden kann. Dadurch werden sie sicherlich helfen, neue Alternativmethoden zu entwickeln, aber sie stellen momentan selbst noch nicht betriebsfertige Alternativen dar.

Einige Probleme im Validierungsprozess

Problem 1: Die Referenz

Abbildung 2 illustriert, was ich das „Validierungsdilemma” nenne: Meist vergleichen wir nicht die Ergebnisse der neuen Tests mit dem, was uns tatsächlich interessiert, d.h. in vivo Daten aus Menschen. Uns stehen solche Daten einfach nicht zur Verfügung. Stattdessen definieren wir den Tierversuch als den (einzigen) Goldstandard. Das bedeutet aber, dass wir von Anfang an die Güte des neuen Tests nur abschätzen können; es ist nicht möglich weiter zu gehen und den Ansatz zu verbessern. Wenn der Tierversuch als richtig angenommen wird, können wir nur eine Sensitivität oder Genauigkeit von weniger als 100% erreichen, und eine Korrelation, die weniger als 1 beträgt. Dies gibt sofort den Eindruck einer geringeren Präzision und Sicherheit.

Das bedeutet (wenn man akzeptiert, dass der Tierversuch tatsächlich nicht perfekt ist), dass alles, was de facto besser ist als der Tierversuch, schlechter abschneidet als er.

Das ist kein wirklich wissenschaftlicher Ansatz, bei dem das zentrale Prinzip ja das Streben nach Besserem sein sollte. Stellen Sie sich vor, dass religiöse Reformatoren von Anfang an die Unfehlbarkeit des Papstes akzeptiert hätten…(um historisch korrekt zu sein: Natürlich wurde das Unfehlbarkeitsprinzip tatsächlich erst vor circa 140 Jahren eingeführt, d.h. lange nach der Reformation). In Zukunft werden wir versuchen, den Begriff „traditioneller Test" statt „Goldstandard" zu benutzen, um die Unsicherheit in Bezug auf den Referenzpunkt auszudrücken.

Es gibt verschiedene Möglichkeiten, das Dilemma zu lösen:

- Wir können den Tierversuch gegen Daten von Menschen validieren (sicherlich nur begrenzt möglich).
- Wir können durch Speziesvergleiche einschätzen, wie gut Wirkungen im Menschen vorhergesagt werden können: Warum sollte die Ratte die Wirkung im Menschen besser vorhersagen als die Maus oder das Meerschwein? In den Fällen, in denen solche Speziesvergleiche durchgeführt worden sind, liegen die Korrelationen meist nur um die 70%.
- Wir können die Reproduzierbarkeit von Tierversuchen durch retrospektive Evaluierung der Variabilität im Test und zwischen verschiedenen Tests abschätzen. Der Unterschied in der Reaktion von Tieren in der gleichen Behandlungsgruppe wird sicherlich kleiner sein als der zwischen Experimenten, die an verschiedenen Tagen oder gar in verschiedenen Labors durchgeführt wurden.
- Wir können eine kombinierte Referenz aus allen verfügbaren Daten erstellen, statt nur mit dem Tierversuch zu vergleichen. Zum Beispiel wird die Klassifizierung einer Chemikalie durch Experten auf Basis aller verfügbaren Daten durchgeführt. Diese können Studien, die nicht nach den Richtlinien durchgeführt wurden, Struktur/Wirkungsbezüge, mechanistische Informationen und Daten von Menschen einschliessen. Diese zusammengestellte Information sollte besser sein als die, die aus einem einzelnen Test gewonnen werden kann.

Trotzdem werden wir in den meisten Fällen weiterhin eine Referenz haben, die zum größten Teil auf Daten aus Tierversuchen beruht, da diese Datensätze von den Regulatoren gefordert werden und zu diesen Zwecken unter den Prinzipien des *Good Laboratory Practice* (GLP) durchgeführt werden. Leider sind eine Menge dieser so genannten hochqualitativen Tierversuchs-Datensätze nicht öffentlich zugänglich, da sie den regulatorischen Behörden vertraulich zur Verfügung gestellt, aber nicht im öffentlichen Bereich publiziert werden. Es ist äusserst wichtig, dass wir aus der Industrie Unterstützung in der Entwicklung von Alternativmethoden bekommen. Diese Information ist auch extrem wichtig für die Entwicklung von Struktur/Wirkungsbeziehungen. Leider ist die Weitergabe von (Roh-)daten meist nicht ausreichend: Auch die Substanz muss zur Verfügung gestellt werden, um die Alternativmethode damit herauszufordern.

In der Praxis existieren hoch-qualitative Datensätze (obwohl vertraulich und ohne Rohdaten) für die Chemikalien, die in den letzten 25 Jahren unter der Direktive für gefährliche Chemikalien in der Datenbank für Neuchemikalien (bisher ca. 4.700 Chemikalien) anmeldet wurden. Dennoch sind nur wenige dieser Chemikalien in ausreichenden Mengen und ausreichender Reinheit käuflich; bei vielen neue Chemikalien wurde die Produktion bereits eingestellt. Hier noch einmal ein Aufruf an die Industrie, Validierung durch die Bereitstellung geeigneter Substanzen zusammen mit ihren Datensätzen zur Verfügung zu stellen.

Hohe Erwartungen werden an die Europäische Partnerschaft für Alternative Ansätze zu Tierversuchen gestellt (EPAA, http://ec.europa.eu/enterprise/epaa/index_en.htm), eine Partnerschaft zwischen Europäischer Kommission und Industrie, die aus einer steigenden Anzahl von einzelnen Firmen (bisher 29) und Handelsverbänden (bisher 7) besteht.

Problem 2: Vorbeugende Toxikologie

Das Prinzip der vorbeugenden Toxikologie ist es, den schlimmsten Fall bezüglich der toxischen Eigenschaften einer Substanz anzunehmen bis es Gegenbeweise gibt. Dieses Konzept entwickelte sich aus der deutschen soziolegalen Tradition der dreissiger Jahre („*Vorsorgeprinzip*"). In 2000 veröffentlichte die Europäische Kommission eine Mitteilung zum Vorsorgeprinzip, in welchem die Vorgehensweise zu dessen Anwendung verabschiedet wurde, wie dies bereits früher in den Verträgen von Maastricht geschehen war. Das Prinzip wurde in viele Bereiche der EU-Politik übertragen, auch in Bereiche, die über die Umweltpolitik hinausgehen, zum Beispiel EU-Lebensmittelrecht und Vorgaben zum Verbraucherschutz, Handel und Forschung und technologische Entwicklung. Eine Arbeitsdefinition und Implementierungsstrategie für den EU Kontext wurde vorgeschlagen (Fisher et al., 2006, freie Übersetzung).

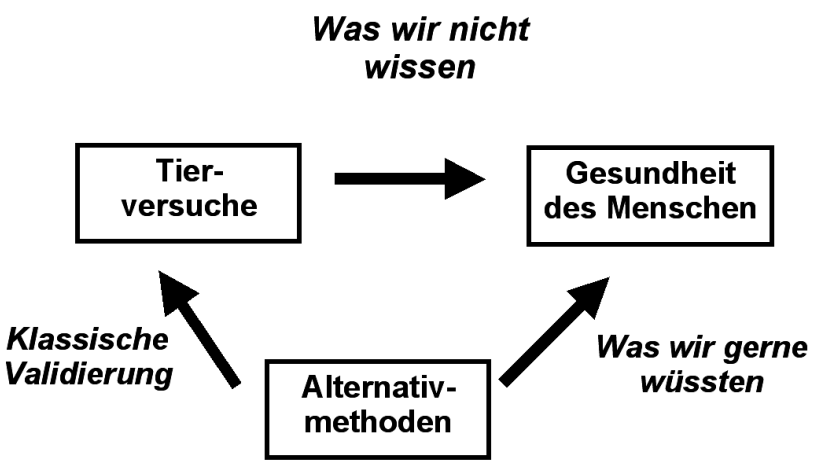


Abb. 2: Das Validierungsdilemma

„Da wo, nach einer Analyse der verfügbaren wissenschaftlichen Information, Gründe zur Besorgnis über mögliche Beeinträchtigungen und wissenschaftliche Unsicherheit fortbestehen, darf auf Basis einer breiten Kosten/Nutzen-Analyse vorläufiges Risikomanagement eingesetzt werden, wobei der menschlichen Gesundheit und der Umwelt Priorität eingeräumt werden, was notwendig ist, um den gewünschten hohen Grad des Schutzes der Gesellschaft und in Verhältnis zu diesem Grad des Schutzes, bis weitere wissenschaftliche Information für eine mehr umfassende Risikoanalyse bereitsteht, ohne warten zu müssen, bis solche möglichen Beeinträchtigungen sich gezeigt haben und das Ausmass ihrer Ernsthaftigkeit vollkommen deutlich ist."

Die Idee ist faszinierend: Lieber ein paar unschuldige Chemikalien opfern als Überraschungen durch Produkte, die auf dem Markt zugelassen wurden, riskieren. Aber, was ist das für eine Art von Referenz für die Validierung, bei der wir eine Inflation von Falsch-positiven haben werden? In manchen Bereichen wie dem Krebs und der Fortpflanzungstoxikologie wurde gezeigt, dass wir es wahrscheinlich mit 10-mal mehr Falsch- als Richtig-positiven zu tun haben (Kirkland et al., 2005; Hoffmann and Hartung, 2005; Bremer et al., submitted; Kirkland et al., 2007). Ob wir uns das für REACH leisten können, d.h. dieses Prinzip auf unsere wertvollsten Chemikalien anwenden, liegt ausserhalb des Rahmens dieser Analyse. Aber wie könnte dieses Prinzip als Standard, als Referenz, für eine Validierung angewendet werden? Es wäre fast unmöglich, einen Test zu entwickeln, der solche Falsch-positive ebenfalls identifiziert. Daher bedeutet so ein vorbeugender Ansatz nicht nur das Ende vieler neu entwickelten Substanzen (und, wenn er in REACH angewendet wird, auch für einige unserer wertvollsten chemischen Produkte). Er stellt auch eine Sackgasse dar, denn er verhindert, dass er selber ersetzt wird, denn jeder neue Ansatz müsste die gleichen Falsch-positiven finden (Hoffmann and Hartung, 2005).

Problem 3: Testrichtlinien statt Standardprotokolle für Tierversuche

Die effektivste Maßnahme, die bisher durchgeführt wurde, um unnötige Tierversuche zu reduzieren, war die Einführung der „gegenseitigen Annahme von Daten" durch die OECD. Das bedeutet, dass Substanzen nur einmal geprüft werden und nicht in jedem einzelnen Land, in dem sie angemeldet werden. Allerdings bedeutet dies meist, dass in der Erstellung der Testrichtlinien Kompromisse getroffen werden müssen, die dazu führen, dass die Tests unpräzise definiert sind und viele Varianten zulassen, damit die verschiedenen Versionen der Tests der verschiedenen Mitgliedsstaaten anerkannt werden können. Das bedeutet, dass die gegenseitig anerkannten Daten aus sehr verschiedenen Versionen der Tests stammen, ohne jeglichen Beweis, dass diese Versionen tatsächlich äquivalent sind. Wie kann so ein diverser Datensatz als Referenz dienen? Wie soll eine Ersatzmethode das Ergebnis einer solchen Sammlung von verschiedenen Tests richtig vorhersagen können? Fast keine dieser Substanzen wird in mehr als einer Variante des Tests eingesetzt. Wer kann dann sagen, welche der Ergebnisse als Referenz für eine Validierung in Frage kommen? Was ursprünglich eine Maßnahme zur Reduktion von unnötigen Tierversuchen darstellte, ist nun ein Hindernis, welches deren endgültige Abschaffung direkt verhindert (Hoffmann and Hartung, 2006).

Problem 4: Standardisierung versus Flexibilität von Alternativmethoden

Das gleiche Problem betrifft Alternativmethoden, bei denen ein sehr spezifisches Protokoll validiert wurde, obwohl vorher und nachher eine Reihe von Varianten verwendet wurden oder werden. Wenn verfügbare Daten für eine retrospektive Analyse zusammengestellt werden sollen, muss ergründet werden, ob die Daten ausreichend ähnlichen Testprotokollen entstammen. Ein Werkzeug aus der klinischen Medizin könnte hier angewandt werden: die Metaanalyse. Dies ist ein Ansatz, bei dem verschiedene klinische Studien kombiniert werden können, um deren Gesamtergebnis zu evaluieren. Allerdings wurden derartige Metaanalysen in unserem Bereich bisher noch nicht durchgeführt. Probleme, die noch geklärt werden müssen, sind, wie Daten, die in der Studie berücksichtigt werden sollen, identifiziert werden sollen, da relevante Daten häufig vertraulich sind, und wie die Qualität der eingeschlossenen Daten kontrolliert werden soll. Was den zweiten Aspekt betrifft, entwickeln wir momentan zusammen mit einem Auftragnehmer die benötigte Qualitätsskala für toxikologische Studien. Wie soll so eine Qualitätsskala aussehen? Zum Beispiel könnte es verschiedene Kategorien geben, die sich von „Einzelbeobachtung in Publikationen ohne *Peer-Review*" bis zum „multizentrischen, verblindeten Ringversuch unter GLP mit unabhängigem Management und unabhängiger Evaluierung" erstrecken würde.

Ein anderer Weg, die Varianten, die für regulatorische Prozesse verwendet werden, zu kontrollieren, wäre, für jeden der validierten *in vitro* Tests einen adäquaten Satz an Leistungsstandards für jeden Anwendungsbereich festzulegen. Dadurch könnte die Äquivalenz einer Testvariante mit der validierten Methode festgestellt werden.

Problem 5: Ersatz eines Tierversuchs durch eine oder durch viele Alternativen

Anders als für akute und topische Toxizität, welche bisher das Gebiet der Entwicklung und Validierung von Alternativmethoden dominiert haben, werden die mehr komplexen Endpunkte nicht durch einzelne Tests ersetzt werden können (abgesehen von einigen Filtertests, durch die bestimmte Substanzen von vornherein aussortiert werden). Stattdessen werden Testkombinationen benötigt. Dieses Konzept nennt sich „intelligente Teststrategie" (ITS), „integriertes Testen", „Teststrategie" oder „Testbatterie" usw. Im Moment fehlen uns die Werkzeuge, solche Teststrategien zusammenzustellen und zu validieren. Sie erfordern sicherlich, dass mehr Substanzen getestet werden, sowohl um die richtige Kombination der Tests in der Teststrategie festzulegen und um die Gesamtstrategie zu validieren. Die Hauptfrage aus meiner momentanen Sicht ist, ob wir alle Einzeltests individuell validieren müssen und können, oder nur die Gesamtstrategie. Der erste Ansatz ist eine Herausforderung in Hinblick auf die Referenz für jeden Teiltest, der letztere Ansatz ist eine Herausforderung in Hinblick auf die Komplexität der

Analyse. Eine Lösung für dieses Dilemma könnte dazwischen liegen: Jeder Teiltest der ITS müsste evaluiert werden in Hinsicht auf Standardisierung und Reproduzierbarkeit (Module 1-4 des modularen Ansatzes), die Relevanz (Module 5 und 6) würde nur für die Gesamtstrategie der ITS evaluiert. Die Diskussion ist noch offen und laufende Integrierte EU Projekte (A-Cute-Tox, ReProTect, Sens-it-iv, OSIRIS und CarcinoGenomics) stehen vor diesen Herausforderungen.

Aber vielleicht wird die Schwierigkeit dieser Herausforderung auch überbewertet: Einige validierte Tests stellen jetzt schon tatsächlich Teststrategien dar – zum Beispiel testet der embryonale Stammzelltest zwei Zelltypen (embryonale Stammzellen und Fibroblasten) mit verschiedenen Endpunkten (kardiale Differenzierung und Zytotoxiziät); die Kombination dieser Test(strategie) hätte möglicherweise mehr analytisch ablaufen können, stellte aber bei der Validierung kein echtes Problem dar.

Problem 6: Fehlende Nachverfolgung nach der Validierung

Die Wissenschaft entwickelt sich immer weiter, und Alternativmethoden werden sich auch nach der Validierung noch Herausforderungen von neuen Ergebnissen und technischen Entwicklungen stellen müssen. Wir müssen für solche Erkenntnisse und Veränderungen offen bleiben; sonst sind wir gefährdet, erneut einen rigiden, traditionellen Ansatz festzulegen.

Ein erster Schritt wäre, das Schicksal validierter Alternativen in der Praxis zu verfolgen. Ein Nutzerforum, Feedback-Mechanismen (z.B. in welchen Fällen die Methode versagte) oder Workshops, bei denen Erfahrungen aus der Praxis (Experimentatoren und Regulatoren) zusammengetragen werden, stellen Chancen dar. Wir müssen offen sein, aufgrund von neuen Erfahrungen die Anwendungsgebiete der Tests sowohl zu begrenzen als auch zu erweitern. Das kann auch einschliessen, dass der Validitätsstatus einer Methode nach einem erneuten *Peer-Review* entzogen wird. Einige Alternativmethoden wurden vor einem Jahrzehnt validiert: Es ist Zeit zu hinterfragen, wo wir heute bezüglich ihrer Implementierung und der Erfahrung aus ihrer Anwendung stehen.

Problem 7: Fehlende Implementierung des relevanten Mechanismus als Kriterium für die Validierung

Während Reproduzierbarkeit/Verlässlichkeit und die Evaluierung der Relevanz gut strukturierte Teile des Validierungsprozesses darstellen, wurde die Evaluierung der wissenschaftlichen Basis der Testmethode bisher noch nicht richtig formalisiert. Sie wird meist als ein Teil der Testdefinition angesehen und die Abwesenheit von offensichtlichen Bedenken wird als Beweis einer gerechtfertigten mechanistischen Basis angenommen.

Allerdings ist es genau diese mechanistische Basis, die Alternativmethoden mit moderner Toxikologie verbindet, welche inzwischen zu grossen Anteilen Mechanismus-basiert ist. Es könnte in einigen Fällen wichtiger sein, dass eine neue Alternativmethode zentrale Mechanismen einer Wirkung auf die Gesundheit oder die Umwelt widerspiegelt, die durch prototypische Chemikalien ausgelöst werden, als zu zeigen, dass ein altmodischer, traditioneller Test mit dem neuen Test nachgestellt werden kann. Dieser Ansatz erscheint vor allem viel versprechend für sich neu entwickelnde Gebiete (Entwicklungsneurotoxikologie, hormonell aktive Schädigung, Immuntoxizität, Atmungsirritation, Atmungssensibilisierung usw.), für die es keine traditionellen Tests gibt. Gleichzeitig stellt dies eine enorme Herausforderung an die Transparenz und die wissenschaftliche Zuverlässigkeit des Validierungsprozesses dar.

Wo stehen wir in Hinsicht auf die Validierung von Alternativmethoden?

Der Prozess der Validierung ist nicht in Stein gemeißelt, sondern kann kontinuierlich optimiert und angepasst werden. Die Definition des modularen Ansatzes (Hartung et al., 2004) stellt nur einen Wegstein dar, der bereits durch Diskussion weiterentwickelt wurde, und mit dem verschiedene Arten der Testvalidierung eingeschlossen werden, wie in Abbildung 3 zusammengefasst (siehe auch: Balls et al., 2006; Hoffmann and Hartung, 2006a).

Wir müssen zunächst unterscheiden, ob die Daten für die Validierung schon erhoben worden sind und retrospektiv analysiert werden sollen, oder ob eine prospektive Studie durchgeführt werden muss. Diese zwei Ansätze schließen sich nicht gegenseitig aus, da Daten, die für die retrospektive Analyse fehlen, durch Daten aus einer prospektiven Studie ergänzt werden können. Während einer laufenden Validierungsstudie kann ein ähnlicher Test aufholen ("*catch-up*") und dann noch bei der Evaluierung mit eingeschlossen werden. Später, wenn der Test bereits validiert worden ist,

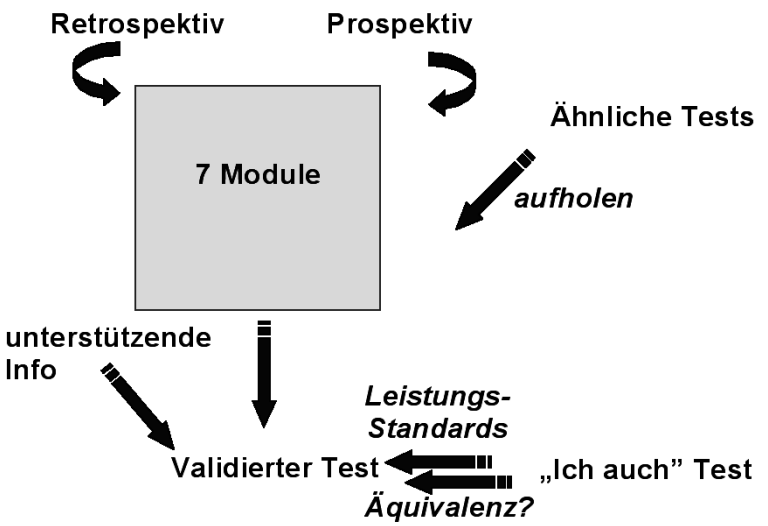


Abb. 3: Validierung und Faktenabwägung

können neue Varianten des gleichen Tests entstehen. Die Frage nach der „Äquivalenz“ dieser „*me-too*“-Entwicklungen (ein Begriff, der der Entwicklung von Generika in der pharmazeutischen Industrie entlehnt ist) stellt sich. Um neue, grosse Studien zu vermeiden, müssen möglichst schon während der ursprünglichen Validierung und dem *Peer-Review* Leistungsstandards festgelegt werden, die bestimmen, was erfüllt werden muss, um Äquivalenz nachzuweisen. ECVAM etabliert gerade das Referenzlabor CORRELATE, um unter anderem solche Evaluierungen von Äquivalenz durchzuführen. Viele dieser Prozesse erfordern eine Abwägung der vorhanden Fakten "Weight-of-evidence" (Balls et al., 2006).

Abbildung 4 fasst einige der offenen Fragen in Hinblick auf die sieben Module (linke Seite) zusammen, die 2004 im Modularen Ansatz definiert wurden. Sie wurden in diesem Artikel angesprochen: Wie können wir die mechanistische Basis erfassen und bewerten? Welche Varianten eines Tests sind äquivalent? Wie können wir die Qualität existierender Daten klassifizieren? Wie führt man eine Metaanalyse durch und welche Daten müssen eingeschlossen werden? Was ist unsere Referenz, wenn kein traditioneller Test zur Verfügung steht? Wie können wir ein Vorhersage-Modell erstellen („*prediction model*“ PM, oder, wie es die OECD nennt, eine Dateninterpretationsprozedur, DIP)? Wie können wir das Anwendungsgebiet eines Tests definieren und später verändern? Wie können wir die Äquivalenz einer Methode, die einer validierten Methode ähnelt, feststellen?

Weitere Diskussion ist notwendig, um die Details auszuarbeiten. Die praktische Arbeit und vielleicht auch die Denkrichtungen werden im Moment sehr stark durch die Bereiche Chemikalien und Kosmetika vorangetrieben; die Potenziale für Pharmazeutika und Grundlagenforschung wurden bereits anderswo diskutiert (Hartung, 2002; Gruber und Hartung, 2004). Diese Gebiete werden aber auch besondere Berücksichtigungen für die Validierung erfordern, wie zum Beispiel:

- Parallele Testung mit den etablierten Tests für Chargenfreigabe/Produktkontrolle
- Produktspezifische Validierung
- Qualitätssicherung von Methoden, die in der Forschung angewendet werden statt definierten Tests (das einzige bisherige Beispiel einer Validierung ist hier die *in vitro* Herstellung von monoklonalen Antikörpern)
- Allgemeine Anliegen der Qualitätskontrolle wie *Good Cell Culture Practice* (Coecke et al., 2005), die inzwischen in verschiedenen Spezialgebieten wie dem der Stammzellen weiterentwickelt wurde
- Möglichkeiten, in der Medikamentenentwicklung präklinische Sicherheitsprüfungen mit den Ergebnissen von Studien in Freiwilligen zu vergleichen (bemerkenswert ist, dass bis zu 30% der Medikamentenkandidaten, die es, nachdem sie die toxikologische Prüfung durchlaufen haben, in Studien mit Freiwilligen schaffen, wegen toxischen Nebenwirkungen aufgegeben werden müssen)
- Methoden für Biologika (vor allem menschliche Proteine und Antikörper gegen menschliche Strukturen), die sich schwer validieren lassen, da relevante Daten aus Tierversuchen fehlen.

Die Veränderungen in der politischen Landschaft (siebte Änderung der Kosmetikdirektive 2003 und REACH 2006) haben ein beispielloses Validierungsprogramm angeworfen: Im Moment sind 187 Testmethoden im ECVAM-Validierungsprozess. Diese sind in sehr unterschiedlichen Stadien des Prozesses (zwischen Reproduzierbarkeitsstudien nach Teststandardisierung bis hin zum endgültigen *Peer-Review*-Prozess bei ESAC, dem wissenschaftlichen Beirat von ECVAM). Ebenso wichtig ist, dass sich ein Ansatz zur Evidenz-basierten Toxikologie abzeichnet (Hoffmann and Hartung, 2006b), in dem Validierung eine Säule der Selbsterneuerung der Toxikologie darstellen soll.

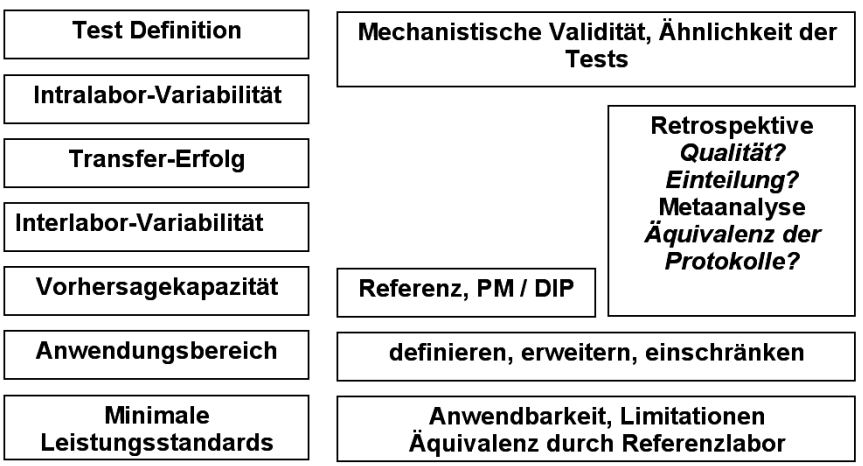


Abb. 4: Offene Fragen des modularen Ansatzes

Literatur

Literaturverzeichnis siehe *“References”* Seite 72f.

Danksagung

Die fortwährende Diskussion und der Einsatz meiner Kollegen in und ausserhalb von ECVAM werden hoch geschätzt. Besonders möchte ich mich bei Sandra Coecke, Christoph Klein und Andrew Burke für die kritische Durchsicht dieses Manuskripts bedanken.

Korrespondenzadresse

Prof. Dr. med. Dr. rer. nat. Thomas Hartung
IHCP-ECVAM
Via E. Fermi 1
21020 Ispra
Italien
E-mail: thomas.hartung@ec.europa.eu