https://mail.gen-ethisches-netzwerk.de/node/2681

Gen-ethischer Informationsdienst

Sicherheitsforschung: Signifikanz und Äquivalenz

Völlige Unbedenklichkeit lässt sich nicht nachweisen

AutorIn
Hans-Peter Piepho

Bei wissenschaftlichen Untersuchungen zu gentechnischen Verfahren und Produkten wird häufig gefordert, dass für mögliche negative Auswirkungen im Vergleich zu Kontrollvarianten ein Signifikanznachweis geführt wird. So haben Kritiker der Séralini-Studie zum Beispiel bemängelt, dass die beobachtete erhöhte Sterblichkeit von Ratten in der dort geprüften Roundup-Variante im Vergleich zur Kontrolle nicht signifikant sei (GID 216, Februar 2013: „Forschung schlägt hohe Wellen“). Diese Kritik ist berechtigt. Die Nicht-Signifikanz der Ergebnisse ist allerdings kein Nachweis für die Unbedenklichkeit, wie in diesem Beitrag herausgestellt wird. Eine völlige Unbedenklichkeit im Sinne einer „Null-Toleranz“ lässt sich grundsätzlich in solchen Untersuchungen nicht nachweisen. Es ist lediglich möglich, eine Äquivalenz („Fast-Gleichheit“) zwischen Kontrolle und einer neuen Behandlung nachzuweisen. Dies erfordert einen ausreichend großen Stichprobenumfang. Dieser Beitrag hat zum Ziel, das statistische Verfahren eines Äquivalenz-Tests und dessen Bedeutung in der Sicherheitsforschung zu erläutern. Birgit Peuker erklärt in ihrem Beitrag „Signifikanzen ohne Unterschiede“ 1 sehr anschaulich die Rolle von statistischen Tests in der Sicherheitsforschung am Beispiel der Séralini-Studie.2 In dieser Studie wurde unter anderem untersucht, ob eine herbizid-tolerante gentechnisch modifizierte Mais-Linie (NK603) eine im Vergleich zur Kontrolle erhöhte Sterblichkeit bei Ratten verursachte. Wie Birgit Peuker berichtet, haben Kritiker der Séralini-Studie bemängelt, dass die beobachteten Todesraten keinem Signifikanztest unterzogen wurden. Einige Kritiker haben gezeigt, dass die beobachtete Erhöhung der Sterblichkeit durch NK603 nicht signifikant ist.3 Hieraus kann geschlossen werden, dass die Versuchsergebnisse nicht geeignet sind, einen negativen Effekt von NK603 nachzuweisen. Damit ist jedoch auch nicht nachgewiesen, dass es einen solchen negativen Effekt nicht gibt.1 Oder mit den Worten von Altman und Bland 4: „Absence of evidence is not evidence of absence“. Denn ein nicht-signifikantes Ergebnis kann immer eine von zwei Ursachen haben. Entweder besteht wirklich kein negativer Effekt, oder aber es besteht in Wahrheit ein Effekt, jedoch war der Stichprobenumfang nicht groß genug, um den bestehenden Effekt nachzuweisen. Ohne weitere Experimente kann nicht geklärt werden, welche dieser beiden Ursachen zutrifft. Strenggenommen lässt sich eine völlige Unbedenklichkeit überhaupt nicht nachweisen. Dazu müsste nämlich gezeigt werden, dass die tatsächliche Differenz in der Sterblichkeit zwischen Behandlung (zum Beispiel die Fütterung mit NK603) und Kontrolle, und damit der Effekt der Behandlung exakt Null ist. Dieser Nachweis ist aber unmöglich, weil es immer eine Zufallsschwankung in den Daten gibt. Dies bedeutet, dass eine Differenz in der Sterblichkeit (oder in anderen Merkmalen) immer nur mit einer gewissen Ungenauigkeit geschätzt werden kann. Diese Ungenauigkeit kann man quantifizieren, indem man ein so genanntes Vertrauensintervall für die Differenz berechnet. Ein solches Intervall überdeckt (enthält, umfasst) die wahre Differenz zwischen Kontrolle und Behandlung mit einer vorgegebenen Wahrscheinlichkeit, beispielsweise 95 Prozent. Je größer der Stichprobenumfang, desto

genauer ist die Schätzung der Differenz, und desto enger ist demzufolge das Vertrauensintervall, das heißt umso näher liegen dessen untere und obere Grenze beieinander. Die untere und obere Grenze werden aber nie identisch sein, was an der unvermeidlichen biologischen Schwankung der Versuchsdaten liegt. Vertrauensintervalle können auch herangezogen werden, um die Signifikanz zu ermitteln. Eine Differenz zwischen Kontrolle und Behandlung ist genau dann signifikant von Null verschieden, wenn das Vertrauensintervall den Wert Null nicht überdeckt. Wird dagegen die Null eingeschlossen, so lassen die Daten keinen Nachweis einer echten Differenz zu, weil die Null dann ein nicht auszuschließender Wert ist. Um umgekehrt nachzuweisen, dass die tatsächliche Differenz exakt Null ist, müssten obere und untere Grenze eines Vertrauensintervalls für die Differenz genau auf den Punkt Null fallen, um für die wahre Differenz alle Werte außer der Null ausschließen zu können. Dies ist aber nicht möglich, weil dies einen unendlich großen Stichprobenumfang erfordern würde. Mit begrenztem Stichprobenumfang lässt sich immer nur nachweisen, dass die wahre Differenz kleiner als ein bestimmter Toleranzwert ist. Das statistische Verfahren hierfür wird *Äquivalenz-Test* genannt.5 Dieses Verfahren unterscheidet sich in seiner Zielstellung grundsätzlich von einem klassischen Signifikanztest. Der klassische Signifikanztest kann zwar im Fall der Signifikanz einen Unterschied zwischen Behandlung und Kontrolle nachweisen. Er lässt im Falle der Nicht-Signifikanz aber keine Schlüsse zu, insbesondere nicht den Schluss der Unbedenklichkeit. Im Unterschied hierzu ist die Zielstellung des Äquivalenztests, einen Signifikanznachweis zu führen, dass Behandlung und Kontrolle sich „kaum“ unterscheiden und daher als „fast gleich“ oder „äquivalent“ zu betrachten sind. Für einen Äquivalenz-Test muss zunächst aus fachwissenschaftlicher Sicht festgelegt werden, welcher Wert für die wahre Differenz noch als tolerabel zu betrachten ist, also bis zu welcher Differenz noch auf eine „Äquivalenz“ („Fast-Gleichheit“) zwischen Kontrolle und Behandlung geschlossen werden kann. Diese Toleranzschwelle wird in der statistischen Literatur gemeinhin mit dem griechischen Buchstaben ? („delta“) bezeichnet. Ein Äquivalenz-Nachweis läuft darauf hinaus, das man zeigen muss, dass ein Vertrauensintervall für die Differenz innerhalb der Toleranzgrenzen ?? und +? liegt (siehe Abbildung 1). Gelingt dies, dann ist damit gezeigt, dass auch die wahre Differenz innerhalb der Toleranzgrenzen ?? und +? liegt, also vom Betrag her nicht größer als der Wert ? ist, was dann als Äquivalenz von Kontrolle und Behandlung interpretiert wird. Dies ist für das oberste Vertrauensintervall in Abbildung 1 der Fall. Wird dagegen eine oder sogar beide dieser Toleranzgrenzen vom Vertrauensintervall überdeckt, wie dies für alle weiteren Vertrauensintervalle in Abbildung 1 der Fall ist, ist der Äquivalenz-Nachweis nicht erbracht; die Daten sind dann nicht ausreichend, um irgendwelche Schlüsse zu ziehen. Um Äquivalenz überhaupt nachweisen zu können, muss der Stichprobenumfang so groß sein, dass das Vertrauensintervall für die Differenz enger wird als das Zweifache von ? (siehe oberstes Intervall in Abbildung 1).

Abbildung 1:
[img_assist|nid=2695|title=|desc=|link=none|align=left|width=500|height=240]

Abbildung in Originalgröße hier.

Eine entsprechende Stichprobenplanung ist daher unabdingbare Voraussetzung für einen Äquivalenz-Test. Um den notwendigen Stichprobenumfang bestimmen zu können, muss neben der Toleranzschwelle noch eine Vorinformation über die zu erwartende Streuung (Varianz) der Daten vorliegen. Die Krux eines Äquivalenz-Tests ist natürlich die Bestimmung des zulässigen Wertes ? für die wahre Differenz von Behandlung und Kontrolle. Wenn eine Null-Toleranz-Politik verfolgt wird, dann kann dies immer nur heißen, dass ?=0 gewählt wird. Damit würde der Toleranzbereich aber kein Bereich mehr sein, sondern er würde auf den Punkt Null zusammenschrumpfen. Dies bedeutet, dass man sich jegliche Experimente sparen kann, weil aufgrund von Versuchsschwankungen nicht zu erwarten ist, dass die in einer Untersuchung geschätzte

Differenz zwischen der Kontrolle und Behandlung exakt den Wert Null annimmt oder das Vertrauensintervall um diese geschätzte Differenz die Breite Null hat. Eine experimentelle Untersuchung macht überhaupt nur dann Sinn, wenn man eine gewisse Differenz zwischen Kontrolle und Behandlung noch für akzeptabel hält. Diese akzeptable Differenz ? muss explizit im Vorhinein festgelegt werden, sonst ist ein Äquivalenz-Test nicht möglich. Was nun ein akzeptabler Wert für ? ist, wird in den meisten Fällen eine umstrittene Frage sein. Fazit: Mit statistischen Mitteln lässt sich eine völlige Unbedenklichkeit im Sinne einer „Null-Toleranz“ einer neuen Behandlung - wie der Fütterung mit der Maislinie NK603 - streng genommen überhaupt nicht nachweisen. Es ist lediglich möglich, einen Äquivalenz-Nachweis zu führen, also zu zeigen, dass die Differenz zwischen Kontrolle und Behandlung eine unter fachwissenschaftlichen Gesichtspunkten festzulegende Toleranzschwelle nicht überschreitet. Hierfür ist ein ausreichender Stichprobenumfang unabdingbar, was bei der Planung entsprechender Studien berücksichtigt werden muss. Entsprechende Verfahren werden derzeit bei der EFSA, der Europäischen Behörde für Lebensmittelsicherheit, diskutiert.

- 1a1bPeuker, B. (2013): Signifikanzen ohne Unterschiede? Gen-ethischer Informationsdienst 216, S. 15-18.
- 2Séralini, G.E. et al. (2012): Long-term toxicity of a Roundup herbicide and a Roundup-tolerant genetically modified maize. Food and Chemical Toxicology 50, S. 4221-4231.
- 3Panchin, A.Y. (2012): Toxicity of roundup-tolerant genetically modified maize is not supported by statistical tests. Letter to the Editor. Food and Chemical Toxicity 53, S. 460.
- 4Altman D.G.; Bland, J.M. (1995): Absence of evidence is not evidence of absence. British Medical Journal 311, S. 485.
- 5Hsu, J.C. (1996): Multiple comparisons. Theory and methods. Chapman & Hall, London, S. 95-96. Da Silva, G.T.; Logan, B.R; Klein, J.P. (2008): Methods for equivalence and noninferiority testing. Biology of Blood and Marrow Transplantation 15 (1 Suppl.): S. 120-127. EFSA Scientific Panel on Genetically Modified Organisms (GMO); Scientific opinion on statistical considerations for the safety evaluation of GMOs, on request of EFSA. EFSA Journal 2009. Im Netz unter www. efsa.europa.eu/en/scdocs/doc/1250.pdf. Beim Äquivalenz-Test wird davon ausgegangen, dass Veränderungen der Behandlung im Vergleich zur Kontrolle unabhängig von der Richtung der Änderung problematisch sind. Wenn nur Veränderungen in eine Richtung problematisch sind, kann man anstelle des Äquivalenz-Tests einen sog. Nicht-Unterlegenheits-Test, einen test of noninferiority durchführen.

Informationen zur Veröffentlichung

Erschienen in:
GID Ausgabe 220 vom Oktober 2013
Seite 28 - 29