https://mail.gen-ethisches-netzwerk.de/node/2487

Gen-ethischer Informationsdienst

Signifikanzen ohne Unterschiede?

Statistische Ergebnisse sind interpretierbar

AutorIn
Birgit Peuker

Die Arbeitsgruppe um Gilles-Eric Séralini hat in ihrer Fütterungsstudie zur Giftigkeit des Herbizides Roundup und von gentechnisch verändertem, gegen Roundup resistenten Mais bei Ratten viele Unterschiede zwischen den einzelnen Untersuchungsgruppen hinsichtlich Sterblichkeit, Ausbildung von Tumoren und 48 weiteren biochemischen Variablen gefunden. Haben sich diese Unterschiede aus Sicht der Statistik nur zufällig ergeben - oder sind sie tatsächlich signifikant?

Einer der wichtigsten Kritikpunkte an Séralinis Artikel, in dem er seine Studienergebnisse veröffentlichte 1, bezieht sich auf die statistische Auswertung seiner Beobachtungen. Im Prinzip wurde angemerkt, dass die von ihm gefundenen Unterschiede zwischen den Untersuchungsgruppen nicht signifikant seien. Tatsächlich hat Séralini in einer Entgegnung auf seine Kritiker zugegeben, dass diese Kritik teilweise berechtigt ist. Seine zentrale statistische Auswertung - eine Diskriminanzanalyse - habe aber zu signifikanten Ergebnissen geführt. Wir wollen uns im Folgenden seine Daten und Auswertungsschritte genauer ansehen, um zu verstehen, um was es in der Diskussion genau geht.

Das Konzept „Signifikanz“

Zunächst müssen wir uns das Konzept „Signifikanz“ genauer ansehen. Signifikanz bedeutet, dass die in einer Untersuchung gefundenen Unterschiede verallgemeinerbar sind, sie sich also nicht zufällig ergeben haben. Demnach müssen zunächst überhaupt erst einmal nennenswerte Unterschiede gefunden werden. Sehen wir uns also an, was für Unterschiede in dieser Studie auftauchen. Dabei müssen wir uns zunächst den Untersuchungsaufbau vergegenwärtigen. Séralini standen 200 Ratten für seine Untersuchung zur Verfügung, die er auf zehn Gruppen aufteilte. Drei Gruppen haben gentechnisch veränderte Organismen (GVO) in unterschiedlicher Dosierung als Futterzusatz bekommen (11 Prozent, 22 Prozent, und 33 Prozent als Beimischung zum normalen Futter). Drei Gruppen haben GVO zusammen mit Roundup ebenso in unterschiedlicher Dosierung und weitere drei Gruppen in unterschiedlicher Dosierung Roundup mit Wasser zu trinken bekommen. Eine weitere, zehnte Gruppe diente als Kontrollgruppe. Sie wurde mit „normaler Nahrung“ gefüttert. Séralini unterschied in diesen einzelnen Gruppen zwischen Männchen und Weibchen. Er hatte also eigentlich 20 Gruppen zu je zehn Tieren.2 Über den Zeitraum von zwei Jahren - der gesamten Lebensdauer des untersuchten Rattenstammes - hat er unter anderem beobachtet, wann die Tiere sterben und was für Tumore sie ausbilden. Weiterhin hat er in regelmäßigen Abständen unterschiedliche biochemische Parameter in 48 Messvariablen bei den einzelnen Tieren erhoben.

Unterschiede bei den Sterbefällen

Die ersten Unterschiede, die in dem Artikel dargestellt werden, beziehen sich auf die Sterbefälle in den verschiedenen, oben genannten Gruppen im Untersuchungszeitraum. In Abbildung 1 sind die Sterbefälle bei den verschiedenen Gruppen dargestellt. Wir sehen sechs Diagramme, links drei Diagramme für die Männchen, rechts für die Weibchen. Nehmen wir als Beispiel das erste Diagramm: Es bezieht sich auf alle männlichen Gruppen, die GVO in unterschiedlicher Dosierung gefüttert bekommen haben. Zusätzlich ist die Kontrollgruppe aufgeführt. In dem Diagramm ist dargestellt, wann wie viele Tiere gestorben sind. Auf der Y-Achse (der senkrechten Linie) ist die Anzahl der Tiere abgetragen. Sie reicht von 0 bis 10, da jede Gruppe 10 Tiere umfasste. Auf der X-Achse (der waagrechten Linie) sind die Tage des Untersuchungszeitraumes von circa 2 Jahren abgetragen, die Achse reicht demnach von 0 bis ungefähr 700 Tage. In dem Diagramm sind die Sterbefälle der vier Gruppen abgetragen: Die dünne Linie (= Graph) repräsentiert die Gruppe, die einen 11-prozentigen Anteil an GVO zu fressen bekommen hat, die mittelstarke Linie die 22-Prozent-Gruppe, die fette Linie die 33-Prozent-Gruppe. Aus der Graphik können wir erkennen, dass das erste Tier nach 100 Tagen gestorben ist. Es handelte sich um eine Ratte aus der 11-Prozent-Gruppe. Die nächsten zwei Tiere derselben Gruppe sterben kurz nacheinander nach dem 400sten und 450sten Tag, anschließend zwei kurz nacheinander nach dem 500sten Tag. In gleicher Weise können wir die anderen Graphen lesen: Aus der 22-Prozent-Gruppe stirbt das erste Tier nach dem 500sten und aus der 33-Prozent-Gruppe nach dem 550sten Tag. Die gepunktete Linie der Kontrollgruppe ist kaum zu erkennen. Der graue Bereich am rechten Rand des Diagramms verdeutlicht den Beginn des Zeitraumes, in welchem die Tiere ohnehin sterben, da ihre Lebensdauer erschöpft ist. Wir können hier also beobachten, dass einige Tiere in der 11-Prozent-Gruppe früher sterben, als die Tiere in der Kontrollgruppe und als die Tiere in den Gruppen, die GVO in höherer Konzentration zu sich nahmen. Wir sind mit dem Lesen des Diagramms aber noch nicht fertig: In der oberen linken Ecke des Diagramms finden wir zusätzlich vier Balken. Das ist ein Histogramm - die Darstellung der (absoluten) Häufigkeit der Sterbefälle zu einem bestimmten Zeitpunkt und zwar kurz bevor das „normale“ Sterben beginnt - bei den Männchen ist das ungefähr der 600ste Tag. Der erste Balken mit der Bezeichnung „0“ stellt die Todesfälle in der Kontrollgruppe dar: Drei Todesfälle sind bei den Ratten mit „normaler“ Fütterung zu diesem Zeitpunkt zu beklagen. Der zweite Balken mit der Bezeichnung 11 stellt die 11-Prozent-Gruppe dar: Hier sind 5 Tiere am Tag 600 bereits verstorben. Wir können also feststellen, es gibt einen Unterschied: Mit drei Todesfällen sind in der Kontrollgruppe weniger Tiere gestorben, als in der Gruppe mit der 11-Prozent-GVO-Fütterung (fünf Todesfälle). Der dritte und vierte Balken verdeutlicht einen weiteren Unterschied: Bei der 22-Prozent-und der 33-Prozent-Fütterung gibt es jeweils zwei Todesfälle weniger als in der Kontrollgruppe. Die Balken enthalten eine zusätzliche Information: Der schwarze Teil verdeutlicht die Fälle, in denen das Tier getötet werden musste, da es zum Beispiel unter großen Tumoren zu sehr litt. Der schraffierte Teil verdeutlicht „spontane“ Todesfälle.

Untersuchung ohne Fragestellung?

Was sagen uns diese Unterschiede, die wir hier beobachtet haben? Wir - als Laien - sehen uns die Zahlen an, weil wir wissen wollen, ob Tiere, die mit GVO gefüttert wurden, schneller sterben als die Tiere der Kontrollgruppe mit normaler Fütterung. Wir interpretieren diese Zahlen also im Hinblick auf eine Fragestellung. Jede statistische Untersuchung macht nur Sinn im Hinblick auf eine zuvor formulierte Fragestellung. In der Einleitung des Artikels können wir eine - spezifischer formulierte - Fragestellung finden: Untersucht werden sollen potentielle, direkte oder indirekte Einflüsse von gentechnisch veränderter Nahrung, von Herbiziden, die mit GVO in Verbindung stehen (hier Glyphosat) oder beidem auf die Gesundheit. Die *European Food Safety Authority* (EFSA) hat trotzdem kritisiert, dass bei Séralinis Studie keine Forschungsfrage angegeben sei, in deren Licht nicht nur das Forschungsdesign, sondern auch die Interpretation der Daten erfolgen könne.3 Die Kritiken beziehen sich vor allem auf die in den Diagrammen in Séralinis Abbildung 1 dargestellten Unterschiede bei den Sterbefällen. Hier lautet der Hauptkritikpunkt, dass die Unterschiede, die wir uns eben gerade angesehen haben, sich auch zufällig hätten ergeben können - mit anderen Worten, die Unterschiede seien nicht signifikant.

Der Zufall mischt mit

Hier spielt der Zufall eine Rolle. Die Untersuchungsobjekte sind nicht vollständig gleich, die jeweiligen individuellen Merkmale, die in den Messwerten erfasst werden, variieren in einem bestimmten Messbereich mehr oder weniger stark. Bezogen auf die Sterbefälle könnten zum Beispiel in die Gruppe der Männchen, die wir oben betrachtet haben, nur Ratten mit einer zufällig etwas schlechteren Kondition eingegangen sein, weswegen sie dann auch früher starben. Die Wahrscheinlichkeit, bei der sich solche Zufälligkeiten auch bei repräsentativen Stichproben ergeben können, kann berechnet beziehungsweise geschätzt werden. Ist die Wahrscheinlichkeit, mit der sich die beobachteten Unterschiede zufällig ergeben können, besonderes gering, spricht man von „signifikanten“ Unterschieden. In Abbildung 1 und auch im Text selbst finden wir keine Angaben zur Signifikanz der beobachteten Unterschiede, obwohl dies eigentlich zur guten wissenschaftlichen Praxis gehört.[4] Wissenschaftler, die sich den Artikel kritisch angesehen haben, haben die Signifikanz nachträglich berechnet.[5] Sie haben dafür den „Fishers-Test auf exakte Signifikanz“ für die Daten aus Abbildung 1 verwendet.[6] Sie sind dabei - ebenso wie wir - von der Annahme ausgegangen, dass Tiere, die GVO als Futter bekommen haben, eine andere Sterblichkeit aufweisen als die Tiere in der Kontrollgruppe. Das heißt, sie haben alle Tiere der sechs GVO-Gruppen (mit und ohne Roundup) je Geschlecht zusammengefasst und mit der Kontrollgruppe verglichen: Von den zehn Männchen der Kontrollgruppe starben drei, von den insgesamt 60 Männchen der GVO-Gruppen starben 19. Das ergibt 3,2 Sterbefälle pro zehn mit GVO gefütterten Tieren, was eine leicht erhöhte Sterblichkeit gegenüber der Kontrollgruppe bedeuten würde. Die Wahrscheinlichkeit nach Fishers-Test, dass sich dieser Unterschied auch zufällig hätte ergeben können, ist P = 0,615. P steht für Wahrscheinlichkeit und kann einen Wert von 0 bis 1 annehmen. Der Wert entspricht einer 61,5-prozentigen Wahrscheinlichkeit, dass sich die leicht erhöhte beobachtbare Sterblichkeit zufällig ergeben hat. Diese Wahrscheinlichkeit ist zu hoch, als dass wir annehmen könnten, dass sich dieses Ergebnis auch bei einer anderen zufälligen Aufteilung der Ratten auf die einzelnen Gruppen ergibt. Bei den Weibchen ist die Wahrscheinlichkeit, dass die Ergebnisse zufällig zustande kamen, geringer. Hier sind die Unterschiede drastischer: 2 zu 10 Todesfälle in der Kontrollgruppe, 29 zu 60 Todesfälle (= 4,8 Todesfälle auf 10 Tiere) in den GVO-Gruppen. Die Wahrscheinlichkeit für ein zufälliges Ergebnis beträgt P=0,09, also 9 Prozent. Das ist schon ziemlich gering, dennoch sind die Ergebnisse nach den gängigen wissenschaftlichen Maßstäben nicht signifikant. In der Forschung hat sich eingebürgert, Ergebnisse erst ab einer bestimmten Schwelle als signifikant zu bezeichnen. Wenn keine großen Konsequenzen aus einer fehlerhaften Schlussfolgerung zu erwarten sind, dann ist man mit einem Signifikanzniveau von 5 Prozent zufrieden. Wenn es dagegen um die Sicherheit auf Leib und Leben geht, wie bei der Sicherheit von Medikamenten und dergleichen, gilt eher 1 Prozent oder gar 0,1 Prozent als Signifikanzniveau.

Ein einzelner signifikanter Fakt reicht nicht aus

Eine weitere Erörterung können wir uns an dieser Stelle ersparen, denn in der Entgegnung auf seine Kritiker hat Séralini bereits eingeräumt, dass die Unterschiede in der Sterblichkeit nicht signifikant sind: „The variability of the mortality can indeed, if interpreted alone, be expected by chance, but in fact the statistics are not powerful enough to conclude that or the contrary. This is why we have described raw data.“[7] Damit weist er auf einen weiteren wichtigen Aspekt hin: Statistische Aussagen müssen nicht nur in einen theoretischen Rahmen, sondern auch durch andere Beobachtungen gestützt sein. Ein einzelner signifikanter Fakt sagt noch nichts: „Statistics do not tell the truth, but may help the understanding.“[8, 9] Die zentrale statistische Auswertung bei Séralini et al. findet sich im zweiten Teil des Artikels. Durchgeführt wurde eine Diskriminanzanalyse, um die Unterschiede zwischen den Gruppen mittels der 48 erhobenen biochemischen Messwerte zu erklären.[10] Vereinfacht ausgedrückt wurde für jede einzelne Gruppe im Vergleich zur Kontrollgruppe ein Modell berechnet, bei der die 48 biochemischen Variablen die Gruppenzugehörigkeit vorhersagen. In diesem Modell wird das Zusammenspiel der 48 Variablen in die Berechnung mit einbezogen. Ausgegangen wird damit davon, dass nicht nur eine Ursache für die Unterschiede zwischen den Gruppen verantwortlich gemacht werden kann, sondern mehrere Faktoren, die sich zudem noch gegenseitig beeinflussen können. Daraus ließe sich dann im Bestfall ableiten, welche Auswirkungen die verschiedenen

Anteile von GVO im Futter auf die Biochemie der Ratten haben. In dem Artikel sind die Ergebnisse für eine Gruppe dargestellt: die Weibchen zum Zeitpunkt „Monat 15" mit der 33-Prozent-GVO-Fütterung (ohne Roundup). Zu diesem Zeitpunkt waren noch die meisten Tiere am Leben. In Graphik 5a sind die Koeffizienten für die einzelnen Messparameter abgetragen. Die Balken geben an, wie stark der Einfluss des jeweiligen Parameters auf die Unterschiede der beobachteten Gruppe (33-Prozent-GVO, weiblich) im Vergleich zur Kontrollgruppe ist. Die darüber gelegten schwarzen Linien repräsentieren jeweils das Konfidenzintervall, das ist das Intervall, innerhalb dessen mit 99-prozentiger Wahrscheinlichkeit der wahre Wert liegt. Wir erinnern uns: Diese wahren Werte können wir nicht wissen, da sich die Unterschiede auch zufällig ergeben können. Hier kann die beobachtete Stärke der Parameter ebenso ein Produkt des Zufalls sein. Wir können in Abbildung 5a in Séralinis Artikel auf den ersten Blick die Signifikanz erkennen: Signifikant sind all jene grauen Balken, bei denen die schwarze Linie für das Konfidenzintervall nicht in den positiven und den negativen Bereich hineinragt. Dies betrifft auf der linken und der rechten Seite je fünf Messwerte (unter anderem Chlor und Natrium; nicht alle Messwerte sind für den biochemischen Laien entschlüsselbar). Mit diesen Werten unterstützt die Arbeitsgruppe um Séralini ihre These gesundheitlicher Auswirkungen von GVO. In der Graphik 5b sind einige dieser Messwerte für die einzelnen Tiere dargestellt.

Höchste Zeit für eine gut geplante Langzeitstudie

Auch an der Abbildung 5a und 5b in Séralinis Artikel ist Kritik geübt worden. So sind in der Graphik nur die Messwerte einer Gruppe dargestellt. Es ist sehr schwierig, im Text eine Begründung für diese Auswahl zu finden. Ebenso entspricht Graphik 5b nicht den üblichen Standards. Dargestellt sind einzelne Messwerte individueller Tiere. Der Graph erweckt den Eindruck, als ob hier ein Wert aufsteigen und danach wieder abfallen würde. Das ist aber nicht der Fall. Einzelne Tiere sind Individuen und stehen nicht in einer Rangfolge, so ist es willkürlich, an welcher Stelle der X-Achse der Messwert abgetragen ist. Ebenso lässt sich bei der Variable „Tier" kein Kontinuum denken, das rechtfertigen könnte, warum hier ein Linie abgetragen ist. Diese Halbheiten und Intransparenzen bei den Graphiken - aber auch im Text - lassen mich schon in Versuchung geraten, den Artikel aus Ärger einfach wegzulegen. Aber man sollte vorsichtig mit einer Pauschalkritik sein. Es empfiehlt sich, weder auf der einen Seite zu behaupten, Gesundheitsrisiken von GVO seien bewiesen, noch den Artikel als nicht wissenschaftlich zu verwerfen. In der wissenschaftlichen Welt gibt es viele Artikel mit ähnlich schlechter Darstellung und ähnlichen uneindeutigen Ergebnissen.[11] Wie auch Séralini in der Replik auf seine Kritiker betonte: Seine Ergebnisse lassen sich sowohl in die eine als auch in die andere Richtung interpretieren. Das heißt, es lässt sich weder ein Risiko aber eben auch keine Sicherheit von Gentechnik ableiten. Wie wäre es damit, den Spuren nachzugehen, auf welche die Studie stieß? Soviel ich verstanden habe, wäre es höchste Zeit für eine gut geplante Langzeit-Studie, einen Versuchsaufbau nach der entsprechenden OECD-Richtlinie, mit transparenter und nachvollziehbarer Darstellung der statistischen Auswertung der Daten. Das wäre allemal besser, als Séralinis Studie in Bausch und Bogen zu verdammen.

- [1]Séralini, Gilles-Eric et al. (2012): Long term toxicity of a Roundup herbicide and a Roundup-tolerant genetically modified maize. In Food and Chemical Toxicology 50 (2012) 4221-4231.
- [2]Die Untersuchungsanordnung wurde ebenfalls kritisiert. Gefordert wurden vor allem mehr Kontrollgruppen. Auf diese Kritikpunkte wird hier nicht weiter eingegangen.
- [3]European Food Safety Authority (EFSA): Review of the Séralini et al. (2012) publication on a 2-year rodent feeding study with glyphosate formulations and GM maize NK603 as published online on 19 September 2012 in Food and Chemical Toxicology. In EFSA Journal 2012;10(10): 2910.
- [4]Vergleiche auch die Kritik beim Bundesamt für Risikobewertung (BfR): Feeding study in rats with genetically modified NK603 maize and with a glyphosate containing formulation (Roundup) published by Séralini et al. (2012). BfR-Opinion 037/2012, 01.10.12.
- [5]Panchin, Alexander Y. (2012): Toxicity of roundup-tolerant genetically modified maize is not supported by statistical tests. Letter to the Editor. In Food and Chemical Toxicology, online November 2012, http://dx.doi.org/10.1016/j.fct.2012.10.039.
- [6]Fishers-Test auf exakte Signifikanz ist ein nichtparametrischer Signifikanztest, der extra für kleine Fallzahlen entwickelt wurde.

- 7Séralini, Gilles-Eric; Mesnage, Robin; Defarge, Nicolas: Answers to critics: Why there is a long term toxicity due to NK603 Roundup-tolerant genetically modified maize and to a Roundup herbicide - Reply to the Letter to the Editor. In Food and Chemical Toxicology, online November 2012, http://dx.doi.org/10.1016/j.fct.2012.11.007.
- 8Ebenda.
- 9In der Diskussion taucht dabei noch eine andere Methode auf. Kritiker (Monsanto) haben darauf hingewiesen, dass insbesondere die Methode von Kaplan-Meier, die speziell dafür entwickelt wurde die Signifikanzen bei der Beobachtung von Sterbefällen zu messen, hätte angewendet werden können. Séralini hat in seiner Antwort auf die Kritiker darauf hingewiesen, dass seine Messergebnisse auch nach dieser Methode nicht signifikant seien. Vgl. Hammond, Bruce; Goldstein, Daniel A.; Saltmiras, David: Letter to the Editor. In Food and Chemical Toxicology, November 2012, http://dx.doi.org/10.1016/j.fct.2012.10.044.
- 10Diese Diskriminanzanalyse heißt OPLS-DA und ist ein multivariantes Verfahren, das ähnlich wie eine Regressionsanalyse über die Auswertung der Varianzen funktioniert. Die unabhängigen Variablen sind hier die 48 biochemischen Variablen. Die abhängige Variable ist in Séralinis Studie eine sogenannte Dummy-Variable: Sie hat nur zwei Werte 0 für „nicht zur Gruppe gehörig" (hier aber nur die Kontrollgruppe) und 1 für„zur Gruppe gehörig". Die Streuungen der einzelnen Variablen werden miteinander verglichen und es wird berechnet, inwiefern die Streuung der unabhängigen Variablen die abhängige Variable erklären kann. An Séralini ist die Kritik geübt worden, dass diese Methode in der Toxikologie nicht üblich sei und von Toxikologen nicht verstanden werden könne (Hammond et al. 2012). Dieser Kritik begegnete Séralini mit dem Verweis auf weitere Anwendungsgebiete dieser Methode. Weiterhin wurde unter anderem von der EFSA kritisert, dass diese spezifische Methode für die Analyse biochemischer Parameter unüblich sei. Das Bundesinstitut für Risikobewertung (BfR) weist darauf hin, dass ein Mittelwertvergleich mit entsprechender Standardabweichung genügt hätte.
- 11Ein Verteidiger von Séralinis Arbeiten hat sich andere Artikel zu Fütterungsstudien angesehen, die auch in dem Journal veröffentlicht worden sind und ausschießlich von Konzernen durchgeführt worden waren. In Bezug auf die Statistik meinte er, dass deren Statistiken genauso aussagekräftig gewesen seien wie bei Séralini. Vielleicht findet sich dereinst eine Öffentlichkeit, die diese Studien skrutinös auseinandernimmt. Heinemann, Jack A.: Food and chemical toxicology. Letter to the Editor. In Food and Chemical Toxicology, online November 2012, http://dx.doi.org/10.1016/j.fct.2012.10.055.

Informationen zur Veröffentlichung

Erschienen in:
GID Ausgabe 216 vom Februar 2013
Seite 15 - 18