Kommentar zur Diskussion

der Arbeiten

Non-thermal DNA breakage by mobile-phone radiation (1800 MHz) in human fibroblasts and in transformed GFSH-R17 rat granulosa cells in vitro

von

E.Diem, C.Schwarz, F.Adlkofer, O.Jahn, H.Rüdiger

Mutation Research 583 (2005), 178-183

und

Radiofrequency electromagnetic fields (UMTS, 1,950 MHz) induce genotoxic effects in vitro in human fibroblasts but not in lymphocytes

von

C.Schwarz, E.Kratochvil, A.Pilger, N.Kuster, F.Adlkofer, H.Rüdiger

Int Arch Occup Environ Health (2008)

1. Einleitung

Die erste Arbeit wird diskutiert in

Vijayalaxmi, J.P. McNamee, M.R. Scafri. Letter to the Editor : Comments on "DNA strand breakes" by Diem et al. [Mutat. Res. 583 (2005) 178-183] and Ivancsits et al. [Mutat. Res. 583 (2005) 184-188]. Mutation Research 603 (2006), 104-106

mit einem Reply to the Letter to the Editor. Rüdiger H.W., E.Kratochvil, A.Pilger Mutation Research 603 (2006), 107-109,

und einem Brief an den Rektor der MUW, July 23, 2007, Not signed, Name known, (der offenbar auch an die Herausgeber von Mutation Research gesendet wurde),

in einem Antwortbrief der Herausgeber an den Verfasser des Briefs Dr.Lerchl

und in einem Antwortbrief von Dr.Lerchl an die Herausgeber.

Die zweite Arbeit wird diskutiert in einem zum Druck angenommenen Brief an den Herausgeber der Zeitschrift Int Arch Occup Environ Health.von Dr.Lerchl

1. Letter to the Editor (Vijayalaxmi et al.)

In diesem Letter wird zunächst methodische Kritik angebracht, die sich generell auch auf andere Arbeiten zu dieser Thematik bezieht („It is important to point out that all of our comments outlined below will also relate to the other previously published reports in which „tail factors" were used to draw conclusions by these investigators").

Allerdings wird auch in dieser Arbeit Kritik an der angewandten Statistik geäußert: " the data ... show negligible standard deviations. It is not clear whether the standard deviations were calculated from a total of 2000 comets (1000 comets from each of duplicate experiments) or from the mean of the two experiments." If the standard deviations were based on 2000 individual comet measurements, then, it is nearly assured that significant differences will be obtained between exposed and sham groups ... . Indeed, it is surprising that such small standard deviations were presented in Diem et al. while in the technical document describing the "tail factor" transformation technique, the standard deviations reported by Diem et al. [2002; J.Toxicol.Environ.Health] were 25% that of the mean."

Die Autoren sind in Ihrem Reply nicht näher auf diese Argumente eingegangen, sie haben vielmehr in einer Tabelle die Daten präsentiert, die einer der Abbildungen der Originalarbeit (Fig. 1) zugrunde liegen.

2. Briefwechsel

2.1 Erster Brief von [redacted]

Der Brief an den Rektor der MUW und die Herausgeber von Mutation Research bezieht sich eben auf diese Daten aus Tabelle 1 des „Reply to the Letter to the Editor". [redacted] verweist auf einige nach seiner Sicht auffällige Zahlenkonstellationen in der Tabelle.

Punkt 1:

Die letzte Ziffer der ersten Spalte („number of cells counted as A") weicht von der erwarteten Gleichverteilung ab („Kolmogorov-Smirnow test reveals a significant ($p<0.04$) deviation from the expected uniform distribution"). Er weist vor allem auf das sehr häufige Auftreten von „2" (14x) im Gegensatz von „5" (1x) hin, allerdings ist nicht klar, wie der χ^2-Test ($p<0.001$) durchgeführt wurde. Offensichtlich sind die 48 Werte in der Spalte 1 nicht unabhängig, da etwa bei den letzten 12 Werten („talkmodulation") die Anzahlen der Zellen in den Replikationen 4 Mal identisch und nur für 2 Versuchbedingungen unterschiedlich sind. Daher sind die Voraussetzungen für die angeführten Tests nicht gegeben. Allerdings ist dann im Gegenzug zu hinterfragen, warum insgesamt die beobachtete Inter-assay-Variabilität kleiner ist als die beobachtete Intra-assay-Variabilität (siehe unten).

Punkt 2:

Beim Zelltyp B entspricht die gefundene Verteilung nicht der Form, wie sie nach einer Poissonverteilung erwartet würde. Die 48 Zahlen in Splate 2 der Tabelle variieren dabei zwischen 35 und 49, sodass die Darstellung der beobachteten Verteilung („observed") durch [redacted] nicht nachvollzogen werden kann.

Punkt 3:

In den Anzahlen der Spalten mit der Zahl von Zellen der Kategorie D und E gibt es Unterschiede in der Verteilung der Zahlen zwischen den Versuchen mit „continuous wave" Exposition zu den drei anderen. Hier kann nicht ausgeschlossen werden, dass solche Irregularitäten bei entsprechend vielfältigen Partitionen der Daten in Subgruppen fast immer gefunden werden können.

Punkt 4:

Bei den Zellzahlen der E-Zellen treten in den 24 Sham-Experimenten 10 mal die Zahl „0" und 14 Mal die „1" auf. Nach der Poissonverteilung „one would expect at least sometimes 2 cells". Dies ist eine nachvollziehbare Kritik (siehe auch unten die Diskussion der niedrigen Variabilität).

Punkt 5:

Tatsächlich ist erstaunlich, dass der Variationskoeffizient des „tail factors" aus den jeweils zwei Wiederholungen in den 24 Versuchsbedingungen durchwegs sehr klein ist („never exceeds 5% and very often below 1%, which is completely incomprehensible for biological data"). Auch im Letter to the Editor (siehe oben) werden unter Hinweis auf Angaben der Autoren in einer Vorpublikation wesentlich höhere Variationskoeffizienten erwartet.

[redacted] schließt daraus: "Taken together, all the arguments listed in this letter strongly suggest that the data in Table 1 of the response letter of Rüdiger et al., and therefore in the paper by Diem et al., were fabricated".

2.2 Antwort der Herausgeber

Die Herausgeber von Mutation Research teilten in Ihrer Antwort [redacted] mit, dass sie den Fall nicht weiter verfolgen werden: "Thus, the Editorial Board and Elsevier fell that we cannot challenge the authors without substantially more evidence in hand – and not just statistical inferences – that data fabrication has occurred."

Als Hauptargument beziehen sie sich auf die in der Arbeit angeführte "blind exposure": „ ... ; and therefore they could not have falsified data to obtain a positive response unless there was collusion by someone at the IT'IS Foundation who informed them of the codes prior to data submission".

2.3 Der zweite Brief von [redacted]

In einer Reaktion auf den Antwortbrief der Herausgeber präsentiert [redacted] weitere Berechnungen. („I can now present undeniable evidence that the data were manipulated / fabricated.")

Er beschränkt sich auf die 12 Kontrollexperimente mit jeweils 2 Wiederholungen aus der Tabelle des „Reply to the Letter to the Editor", in denen jeweils nur „Sham-exposure" zur Anwendung kam. Zunächst weist er auf die sehr geringe Variabilität zwischen den 12 Mittelwerten aus den jeweils zwei Wiederholungen hin (CV: 1.26 %), die geringer ist als die durchschnittliche Variation innerhalb der Wiederholungen (Mittelwert der 12 CVs für die Doppelmessungen: 2.07%). Würde die Streuung zwischen den Doppelbestimmungen genau so groß sein wie die Streuung der Bestimmungen über die unterschiedlichen Experimente, so sollte die Streuung der Mittelwerte aus den beiden Wiederholungen um den Faktor $1/\sqrt{2}$ kleiner sein als die zwischen den Wiederholungen. Die beobachtete Variabilität der Mittelwerte ist immer noch etwas kleiner als es diesem Faktor entspräche. Da jedoch erwartet wird, dass die Streuung zwischen den Experimenten deutlich größer ist als die Streuung innerhalb der Wiederholungen, wird das Datenmaterial angezweifelt („...; an inter-assay variation smaller than the intra-assay variation is obviously nonsense").

Zum Nachweis der Unplausibilität des Datensatzes wurden Zellen der Sham-Experimente aus der Tabelle in einen Pool zusammengeworfen („The number of cells of the 5 categories A-E from Table 1 of the response letter were used to create a „random cell suspension" of 10,000 „cells""). Allerdings sollte durch ein Poolen aller dafür in Frage kommenden Daten 24x500= 12,0000 Zellen resultieren, sodass hier keine Klarheit über die tatsächlich angewandte Methode des Poolens besteht. Dies ist ein kritischer Punkt, da ein Zusammenwerfen aller 12,000 Zellen aus Tabelle 1 das natürliche Vorgehen gewesen wäre. Damit wäre ein Pool entstanden, der exakt die Verteilung aller im Originalexperiment gefundenen Zellen repräsentiert hätte. Aus dem (wie auch immer geschaffenen) 10,000-Pool wurden dann 24 Mal je 500 Zellen zufällig ausgewählt, um die 24 Sham-Experimente zu „simulieren". Die 24 simulierten Experimente wurden dann in gleicher Weise ausgewertet wie die 24 Originalexperimente, was zu 24 simulierten Werten des Tail-factors führte. Dabei zeigte sich, dass in der Simulation beide Variabilitäten (intra- und inter- assay) deutlich höher lagen als in den Originalexperimenten, während der durchschnittliche Tail-factor über alle Experimente keinen nennenswerten Unterschied zwischen Simulation und Realität zeigt. Letzteres ist aufgrund des Vorgehens der mehrmaligen Entnahme einer zufälligen Stichprobe aus dem „gepoolten" Originalexperiment zu erwarten und wird als Bestätigung für die Richtigkeit der Simulation gewertet („ ... shows that the simulation worked as expected"). Auch für die Anzahl der Zellen in den verschiedenen Kategorien zeigt die Simulation deutlich größere Variationen als das Originalexperiment.

Die Simulation beruht auf der Gesamtvariabilität der Originalexperimente, da die Wiederholungen bei den 12 Experimenten ebenfalls aus dem gesamten Pool simuliert werden. Dies kann im vorliegenden Fall dadurch gerechtfertigt werden, dass bei den tatsächlichen Beobachtungen unter den 12 Sham-Experimenten in der Originaltabelle die für den Tail-factor gefundene Inter-assay-Variabilität nicht größer ist als die Intra-assay-Variabilität. Wenn die im Originalexperiment gefundenen

Variabilitäten kritisch eingeschätzt werden sollen, dann sollte jedoch die Verteilung der Variabilitäten durch mehrfache Simulation des gesamten Versuchs geschätzt werden. Eine einzige Simulation des Versuchs ist für eine solche Bewertung nicht ausreichend.

3. Letter to the Editor [redacted]

Dieser jüngst von [redacted] verfasste Brief an die Herausgeber wurde zum Druck angenommen und bezieht sich auf die zweite Arbeit (Schwarz et al.).

„Low standard deviation"

Wieder wird als erster Kritikpunkt die niedrige Variabilität der Ergebnisse genannt. Dabei ist allerdings das Argument der höheren Variationskoeffizienten bei den Anzahlen der E-Zellen als bei den Werten des berechneten „comet tail factors" (CTF) nicht ganz schlüssig ist. Durch eine gewichtete Kumulation der Zellzahlen kann es sehr wohl zu einer Verringerung der Standardabweichung kommen. Allerdings ist dabei zu bedenken dass gerade die Anzahlen der E-Zellen das höchste Gewicht haben („67.5% of fragmented DNA") und daher wesentlich zur Variabilität des CTF beitragen. Die Argumentation, dass für die Standardabweichungen des CTF die Anzahl der Zellen keine Rolle spielt, ist nicht korrekt. Die gewichteten Summe von Anteilen aus 500 Zellen (und genau so ist der CTF definiert) hat eine Standardabweichung die um einen Faktor der Größenordnung $\sqrt{500/50} = \sqrt{10} \sim 3.16$ kleiner sein sollte als die Standardabweichung der mit gleichen Gewichten gewonnenen Werte auf der Basis von nur 50 Zellen. Die beobachteten Unterschiede zwischen den Variationskoeffizienten in der Arbeit von Schwarz et al. und den Variationskoeffizienten aus den unabhängig durchgeführten Experimenten an jeweils 50 untersuchten Zellen (abschätzbar aus den grafischen Darstellungen des S.E.M. in Speit et al. [Mutation Research 626 (2007) 42-47]) sind allerdings immer noch sehr viel größer als es diesem Faktor entsprechen würde.

Dass bei einem solchen Experiment wiederholt derartig geringe Variabilitäten auftreten (bei Scheinbehandlungen und negativen Kontrollen, im zeitlichen Verlauf und über die verschiedenen Expositionen), bleibt also weiter ein ernstes Problem (siehe Abschnitt 4).

„Calculation errors and statistical analyses"

Wie richtig bemerkt, sollten die mittleren Anzahlen der Zelltypen in Tabelle 2 der Arbeit bis auf Rundungsfehler in Summe 500 ergeben. Hier treten Diskrepanzen auf, die aufgrund des Designs nicht erklärt werden können.

Wie auch richtig bemerkt, ist der Rangsummentests für den Vergleich von zwei Stichproben auf einem zweiseitigen Signifikanzniveau von 0.05 bei nur drei Beobachtungen pro Stichprobe nicht anwendbar. Offenbar wollte man sich nicht einer möglichen Kritik and der Voraussetzung für den t-Test (Normalverteilungen mit gleicher Variabilität in den zu vergleichenden Gruppen) aussetzen und hat daher

jeweils Scheinbehandlung und Negativkontrolle in eine Gruppe zusammengefasst und diese mit den jeweiligen Expositionsgruppen im Rangsummentest verglichen. Dies war offensichtlich eine „post hoc" Entscheidung, da kein Unterschied zwischen den beiden „unbehandelten" Gruppen beobachtet worden war. Dieser Vorgang hatte den weiteren Vorteil, dass vermieden wurde, für einen t-Test zwischen zwei Gruppen mit jeweils nur drei Werten extrem kleine p-Werte berichten zu müssen. Schon bei der niedrigsten Dosis in Tabelle 2 hat der Abstand der Mittelwerte zwischen „Exposed" und „Sham" die Größenordnung von 19 (!) Standardabweichungen der Einzelwerte innerhalb der Gruppen. Diese relativen Effekte (in Einheiten der biologischen Variabilität) sind von einer Größe, wie sie im Kontrollexperiment von Speit et al. (nach einer groben Abschätzung aus den Abbildungen) nicht einmal nach einer Exposition gegenüber zwei Gy ^{137}CS Gammastrahlen der Intensität 4GY/min beobachtet werden konnten.

4. Ein Kommentar

Unabhängig von den oben bereits angesprochenen Details aus den Diskussionen über die erste Arbeit zeigen die Daten der Tabelle 1 im „Reply to the Letter to the Editor" aus statistischer Sicht sehr auffällig niedrige Variabilitäten.

So variieren in der Tabelle die Anzahlen der A-Zellen über die 24 Sham-Experimente nur von 443-453 (Anteile: 0.886-0.906), der B-Zellen von 35-46 (Anteile: 0.07-0.092) und der C-Zellen von 8-12 (Anteile: 0.016-0,024). Legt man in einem simplen Modell die Variabilitäten der Multinomialverteilung zugrunde (und nimmt vereinfachend an an, dass die Experimente keine zusätzliche über die Wiederholungsvariabilität hinausgehende Variation generieren), so würden bei „wahren" Anteilen von 0.9, 0.08 und 0.02 die Standardabweichungen der beobachteten Anzahlen in Stichproben vom Umfang 500 die Werte 6.7 ($=\sqrt{500x0.9x(1-0.9)}$), 6.1 und 3.1 annehmen. Die gefundenen Zellzahlen im Originalversuch fallen also alle jeweils in einen Bereich, dessen Gesamtlänge immer kleiner als 2 Mal die einfache Standardabweichungen ist. Die geringe Variabilität findet sich entsprechend auch bei den Abweichungen der in den 24 Experimenten beobachteten Anzahlen zu den jeweiligen Gesamtmittelwerten aus allen 24 Experimente (A: 449.63; B: 38.79; C: 10.08). In Relation zu den nach der Multinomialverteilung anzunehmenden Standardabweichungen sind diese im Originalexperiment gefundenen Abweichungen gering. Man würde sich bei 24 Wiederholungen und Werten für jeweils 3 Klassen von Zellen mit nennenswerten Anzahlen (A, B, und C) nach den Regeln der Stochastik erwarten, dass sich die 72 Werte nicht derart stark in einem engen Bereich zusammenballen. Auch bei den D-und E-Zellen ist die Variabilität gering. So wurde etwa, wie schon von Dr.Lerchl darauf hingewiesen, bei den E-Zellen 10 Mal „0", 14 Mal „1" und niemals „2" oder höher gezählt.

Zunächst kann nicht generell ausgeschlossen werden, dass in einem Experiment eine auffällige Datenkonstellation durch Zufall beobachtet wird. Jüngst ist jedoch eine zweite Arbeit erschienen (Schwarz et al. 2008), obwohl es in der Zwischenzeit nicht gelungen war, die Effekte und niedrigen Variabilitäten in einem unabhängigen Versuch zu reproduzieren (Speit et al., Mutation Research 626 (2007) 42-47: "Because of the ongoing discussion on the biological significance of the effects

observed, it was the aim of the present study to independently repeat the results using the same cells, the same equipment and the same exposure conditions. ... For both tests, clearly negative tests were obtained in independently repeated experiments. ... The reasons for the difference between the results reported by the REFLEX project and our experiment remain unclear.").

Ohne diesmal auf die Originaldaten der Zellzahlen zurückgreifen zu können, zeigen die Daten in der neuen Publikation wieder ein für derartige Versuche unplausibles Muster extrem niedriger Variabilität und höchster Reproduzierbarkeit über „Dosis" und Zeit in Versuchgruppen mit jeweils nur drei Wiederholungen. Nimmt man z.B die Zahl der A-Zellen in den 10 Versuchsreihen ohne echte Exposition aus Tabelle 2 dieser Arbeit („Sham", „Negative control") und bildet den Durchschnitt der angegebenen Mittelwerte aus den jeweils drei Wiederholungen über alle 10 Versuchsreihen, so erhält man ein Anteil von 0.89 über alle 15000 (=3 x 10 x 500) Zellen. Dieser Anteil ist dem Wert 0.9 für A-Zellen aus den analogen Versuchen ohne Exposition in Tabelle 1 des „Reply to the Letter to the Editor" zur ersten Arbeit sehr ähnlich ist (siehe oben). Man kann nun wieder den Ergebnissen in dieser Tabelle 2 das einfache Modell unabhängiger Stichproben vom Umfang 500 aus einer Multinomialverteilung zugrunde legen, diesmal mit einer nur geringfügig kleineren Wahrscheinlichkeit von 0.89 für Zellen der Art A. Dann sollte die Standardabweichung der beobachten Anzahlen von A-Zellen in den Einzelversuchen aus je 500 Zellen den Wert $\sqrt{500x0.89x(1-0.89)} = 7.0$ annehmen. Nun ist es einfach, diesen erwarteten Wert mit den aus jeweils 3 Wiederholungen berechneten Standardabweichungen in den 10 Versuchsreihen ohne echte Exposition (0.05, 0.1, 0.5, 1.0 und 2.0 W/kg, jeweils für „Sham" und „Negative control") aus Tabelle 2 zu vergleichen. Man sieht, dass diese 10 Standardabweichungen zwischen 2.29 und 5.55 schwanken, wobei 9 der zehn den Wert 3.8 nicht überschreiten und 7 nicht den Wert 3.0. Wieder kommt einer solchen Konstellation, bei der in 10 unabhängigen Stichproben die gefundenen Standardabweichungen alle deutlich unter dem nach dem Modell erwarteten Wert liegen, eine geringe Wahrscheinlichkeit zu (trotz der Unsymmetrie der Verteilung um den Erwartungswert). Das gleiche Bild zeigt sich bei den B-Zellen, wo die beobachteten 10 Standardabweichungen zwischen 1.51 und 4.51 schwanken und wieder alle unter den (bei einem theoretischen Anteil von 0.08 an B-Zellen) erwarteten Wert 6.1 fallen. Noch extremer ist das Bild bei den C-Zellen. Die 10 gefundenen Standardabweichungen liegen zwischen 0.62 und 1.78, wobei alle nicht einmal halb so groß sind wie der erwartete Wert von 3.8 (unter der Annahme eines theoretischen Anteils von 0.03 an C-Zellen). Hier treten also wieder, wie in der ersten Arbeit in den Versuchen ohne echte Exposition, bei den A, B und C-Zellen höchst unplausible Datenkonstellationen auf, denen unter einfachen statistischen Voraussetzungen eine extrem geringe Wahrscheinlichkeit zukommt.

Eine genaue Bestimmung der Wahrscheinlichkeiten für solche Auffälligkeiten in den zwei diskutierten Arbeiten ist nicht möglich, da bei solchen Problemen die Fragen im Allgemeinen erst nach Betrachtung der Ergebnisse gestellt werden („data driven analysis"): Bei entsprechend langer Suche wird immer wieder die eine oder andere auffällige Konstellation zu finden sein, die durch Zufall entstanden sein kann. Es soll daher auch hier kein Versuch einer Abschätzung dieser Wahrscheinlichkeiten unternommen werden.

Allerdings muss davon ausgegangen werden, dass die Wahrscheinlichkeit dafür, dass sich derartige unplausible Konstellationen der gleichen Art in Serie durch Zufall ergeben, extrem gering ist.

5. Zusammenfassung

Extreme Ergebnisse können nicht völlig ausgeschlossen werden, und statistische Überlegungen alleine sind, wie in der Stellungnahme der Herausgeber von Mutation Research im Antwortbrief an [redacted] angeführt, nicht ausreichend, wenn Probleme der vorliegenden Art befriedigend gelöst werden sollen (siehe 2.2).

Dabei ist zunächst zu würdigen, dass sich die Gruppe der Herausforderung einer unabhängigen Validierung ihrer Ergebnisse durch eine externe Forschergruppe gestellt hat. Obwohl im Detail nicht genau nachvollziehbar, haben die Autoren aus den beiden Ländern Österreich und Schweiz auch beschrieben, dass ihre Versuche unter Blindbedingungen durchgeführt wurden (z.B. Schwarz et el: „To enable blind experimentation, a computer randomly determined which of the two waveguides was exposed. This setting, which could neither be controlled nor ascertained by he experimenter, was stored in an encoded file and uncovered by the ITIS foundation in Zurich via e-mail in exchange with the transmission of results.").

Trotzdem liegt Evidenz vor, dass ein Versuch einer unabhängigen Forschergruppe völlig fehlgeschlagen ist, die wiederholt von der Gruppe publizierten, ähnlichen und schon im Einzelfall unplausiblen sowie wenig wahrscheinlichen Datenmuster unabhängig zu reproduzieren. Daher müssen an der Valididät der Ergebnisse in den beiden diskutierten Arbeiten fundamentale Zweifel angemeldet werden. Es kann auch nicht gelten, dass die Kritik ausschließlich durch [redacted] betrieben wird (aus welchen Motiven auch immer), da schon als Reaktion auf die erste Arbeit frühzeitig ein „Letter to the Editor" aus einer internationalen Autorengruppe erschienen ist, in dem Überraschung über die im Vergleich zu früheren Experimenten erzielte niedrige Variabilität ausgedrückt wurde.

[redacted] Wien, 6.5.2008

[redacted]

[redacted]