Regression zum Mittelwert

Markus Gnädinger[a], Peter Kleist[b]

[a] Institut für Hausarztmedizin der Universität Zürich
[b] GlaxoSmithKline AG, Münchenbuchsee

Welche Aussage zur Wirkung von Medikamenten ist unmissverständlich?

a) Das Medikament A reduziert bei Patienten mit COPD die Anzahl von Exacerbationen um 50%.
b) Das Medikament B senkt den systolischen Blutdruck um 12 mm Hg.
c) Das Medikament C reduziert die Dauer und die Stärke von Migräneanfällen um je 30%.
d) Das Medikament D reduziert bei Patienten mit MS die Anzahl von Schüben um 20% gegenüber der Häufigkeit unter Plazebo (absolute Risikoreduktion 2% in fünf Jahren).

(Antwort d)

Hätten Sie's gewusst?

Der Term «Regression towards the mean» (später auch: «Regression to the mean») wurde von Sir Francis Galton Ende des 19. Jahrhunderts gemünzt [1]. Anhand von Messungen der Körpergrösse und des Körpergewichts von Kindern besonders grosser Eltern stellte er fest, dass die Kinder kleiner waren als aufgrund der elterlichen Werte zu erwarten gewesen wäre; die Messwerte hatten sich also in Richtung des Durchschnitts, das heisst «zur Mitte», bewegt.

Das Phänomen der Regression zum Mittelwert kommt in vielen Bereichen des praktischen Lebens vor und kann uns in die Irre leiten. Anfällig für dieses Phänomen sind zum Beispiel Unfallstatistiken (die belegen sollen, dass Massnahmen an Verkehrskreuzungen mit hoher Unfallrate tatsächlich etwas genützt haben).

Im Zusammenhang mit wissenschaftlichen Studien bezieht sich der Ausdruck «Regression zum Mittelwert» (RZM) auf die Tatsache, dass sich Krankheitsmerkmale von in Studien eingeschlossenen Versuchspersonen bei Verlaufsmessungen auch in der Plazebogruppe regelhaft bessern. Dieses statistische Phänomen führt dazu, dass in Studien die Wirksamkeit von medizinischen Interventionen überschätzt wird, wenn isoliert nur die Veränderung im Verum-Arm betrachtet wird. Die Testfragen im Kasten weisen darauf hin, dass die Information aus einer korrekt durchgeführten Studie irreführend wiedergegeben werden kann, wenn der Zusatz *«im Vergleich zu Plazebo»* fehlt.

Die RZM kommt bei allen Verlaufsuntersuchungen vor, sofern erstens das Einschlusskriterium und der Erfolgsparameter auf der gleichen Messgrösse beruhen (z.B. Blutdruckwerte) und zweitens dieser Parameter eine gewisse Variabilität aufweist, beispielsweise durch Spontanschwankungen oder Messungenauigkeiten. Sicher ist sinnvoll, sich bei klinischen Studien auf jene Personen zu konzentrieren, die eine mittlere bis starke Schwere der zu behandelnden Krankheit aufweisen; dies reduziert die Zahl der einzuschliessenden Personen und somit die Kosten für die Studie. Trotzdem muss man sich im Klaren sein, dass der Preis dafür eben die Auslösung eines RZM-Phänomens ist.

Gedankenexperiment mit Würfeln

Damit dem Leser bewusst wird, dass es sich hier nicht um ein psychologisches Phänomen, sondern um ein reines Zahlenproblem handelt, haben wir – anstelle von klinischen Beispielen – ein Gedankenexperiment mit Würfeln gewählt. So kann die Leserschaft, sofern sie den Aufwand nicht scheut, unsere Behauptungen selbst nachprüfen, ohne Patienten oder gar eine Ethikkommission zu behelligen …

Statt eines einzigen Würfels nehmen wir drei und werten die Summe der Augenzahl aus. Und – die Menschen sind ja auch verschieden – statt einer Sorte Würfel nehmen wir sechs (Typ A: 1–6 Augen; B: 2–7; … bis F: 6–11). Werfen wir drei Würfel desselben Typs, können Augensummen von $3 \times 1 = 3$ bis $3 \times 11 = 33$ beobachtet werden. Die Abbildung 1 zeigt die erzielte Verteilung der gewürfelten Augenzahlen, die einer Normalverteilung recht nahe kommt. Aus dieser Grundgesamtheit wollen wir die Würfe selektionieren, bei denen eine Zahl über 21 Augen erzielt worden ist. Wir schliessen somit von den ursprünglichen 1296 Varianten 397, das heisst knapp ein Drittel, ein. Die Abbildung 2A zeigt die neue Verteilung. In dieser kommen keine Würfel des Typs A oder B vor, da das Kriterium von 22 Punkten durch diese nicht erreicht werden kann. Jede Kategorie umfasste ursprünglich 216 Beobachtungen; eingeschlossen wurden C: 10 (5%); D: 56 (29%); E: 135 (63%) und F: 196 (91%). Wenn wir annehmen, dass nur die Würfel des Typs E und F hätten eingeschlossen werden sollen, so lässt sich feststellen, dass 77% der Zielgruppe rekrutiert worden sind, während 20% der Rekrutierten eigentlich nicht in die Zielgruppe gehört hätten.

Stellen wir den 397 Beobachtungen je eine Verlaufsbeobachtung entgegen, indem wir nochmals drei Würfel des betreffenden Typs werfen, resultiert eine neue Verteilung (Abb. 2B). Der Mittelwert ist dabei von 25,0 auf 23,4 und somit deutlich in Richtung Mittelwert der ursprünglichen Verteilung (18,0) gerückt. Die neue Verteilung ist von der alten höchst signifikant verschieden ($p < 0{,}001$).

Die Autoren haben keine finanziellen oder persönlichen Verbindungen im Zusammenhang mit diesem Beitrag deklariert.

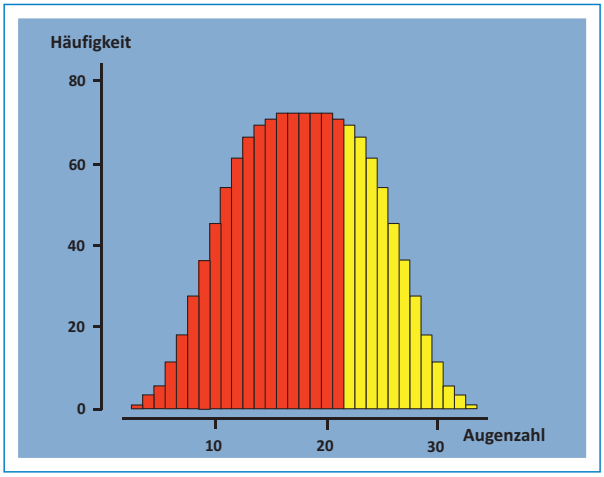

Abbildung 1
Häufigkeit der Augenzahlen dreier Würfel (Typ A, «normal»: 1–6; B: 2–7; C: 3–8; D: 4–9; E: 5–10; F: 6–11 Augen). Es sind bei jeweils drei Würfeln desselben Typs insgesamt 1296 Permutationen möglich. Es resultiert eine Verteilung mit dem Mittelwert 18,0 und einer Standardabweichung von 5,9. In unserem Beispiel wird ab der Zahl 22 etwa ein Drittel der Population mit der höchsten Punktezahl selektiert (gelb hervorgehoben).

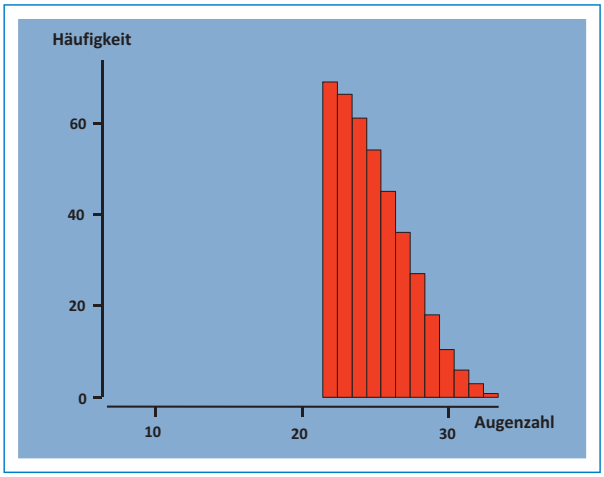

Abbildung 2A
Augenverteilung bei Studieneinschluss (Variantenzahl: 12; Mittelwert: 25,0; Standardabweichung 2,4; N = 397).

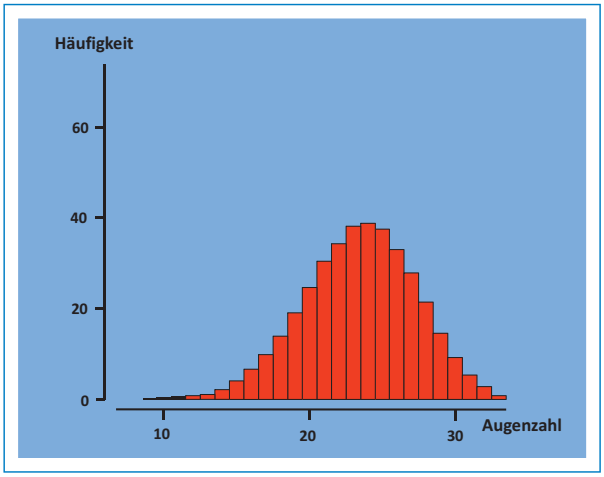

Abbildung 2B
Augenverteilung bei Nachuntersuchung (Variantenzahl 25; Mittelwert 23,4; Standardabweichung 3,8; N = 397), t = 31,4, p <0,001.[1]

Wollen wir die Effektstärke beurteilen, können wir als einfachstes Mass Cohens d verwenden. In unserem Beispiel resultiert ein Wert von 0,5. Dies entspricht einer *mittleren* Effektstärke (ohne dass eigentlich ein Effekt da war!).
Somit können wir folgern, dass, sofern eine Variabilität der Messerwerte vorhanden ist, jede wissenschaftliche Studie, die ein bestimmtes Merkmal zugleich selektioniert und durch eine Intervention verändern will, einem RZM-Effekt unterliegt. Bei adäquater Gruppengrösse weist die Verlaufsbeobachtung zwingend eine statistisch signifikante Differenz zum Vorwert auf.
Unsere Darlegungen gelten demzufolge für jegliche vorstellbaren klinischen Verlaufsparameter, die mit einer therapeutischen Intervention verbessert werden sollten, wie zum Beispiel Werte der Hamilton-Depressionsskala, Quadratzentimeter psoriatischer Plaques, Reizdarm-Symptomskalen, Anzahl neuer Brustwirbelköperfrakturen etc.

Kommt die RZM in der täglichen Praxis auch vor?

Würden Sie bei einen Patienten, der normalerweise einen Blutdruck von 140/90 mm Hg hat, eher eine Intensivierung der antihypertensiven Therapie erwägen, wenn er mit einem Blutdruck von 160/100 mm Hg in die Praxis kommt? Falls «Ja», dann unterliegen auch Sie dem RZM-Effekt. Wenn er das nächste Mal kommt und mit der neuen Medikamentenkombination wieder einen Blutdruck von 140/90 mm Hg hat, ist es dann das neue Medikament, das wirkt? Wer weiss ...!
Welche Massregeln zu berücksichtigen sind, um dem RZM-Effekt in möglichst geringem Mass zu unterliegen, fassen wir in Tabelle 1 zusammen. Einerseits kann die Präzision des selektierten Merkmals verbessert werden, wenn mehrere Messungen gemacht werden, bevor ein Patient in eine Studie eingeschlossen wird (in unserem Beispiel hätten zehn statt dreier Würfe die Typisierung der Würfelklasse verbessert). Zudem kann, sofern dies möglich ist, zu Beginn eine Leerphase ohne Intervention, das heisst eine «placebo run-in phase», gemacht werden; dann werden die Messwerte nach Intervention nicht mehr mit den Werten zum Zeitpunkt des Einschlusses in die Studie, sondern mit jenen am Ende der

[1] Die Abbildung 2B ist aufgrund gewichteter Daten errechnet, so dass am linken Ende auch unter-ganzzahlige Säulen erscheinen. Aus der rechts-schrägen (2A) ist eine links-schräge Verteilung geworden. Tatsächlich verfehlen die Verteilungen in der Abbildung 2A deutlich und in 2B leichtgradig eine Normalverteilung; das richtige Mass für die zentrale Tendenz wäre demzufolge der Median (A: 25; B: 24) und für die Streuung die Interquartilsdistanz (A: 4; B: 5). Da diese nichtparametrischen Masse aber weniger anschaulich sind und da die RZM für nichtnormalverteilte Daten genauso bedeutsam ist wie für normalverteilte, lassen wir es dabei bewenden. Unser einfaches Würfelmodell war zudem nicht in der Lage, gepaarte Daten zu erzeugen. Der richtige Test, um eine Verschiedenheit der zentralen Tendenz der beiden Gruppen nachzuweisen, wäre demzufolge der gepaarte t-Test gewesen oder sein nichtparametrisches Gegenstück, der Wilcoxon-signed-rank-Test. Aufgrund des exzessiv hohen t-Wertes wären zweifellos auch die alternativen Tests höchst signifikant ausgefallen.

Tabelle 1
Massnahmen zur Reduktion des RZM-Effekts in wissenschaftlichen Studien [2].

Beim Studiendesign
– Randomisierte Studiengruppenzuteilung
– Mehrfache Messung der untersuchten Variable vor Studieneinschluss
– Leerphase zu Beginn der Studie vor der Intervention («placebo run-in phase»)
Bei der Datenauswertung
– Graphische Darstellung (Änderung vs. Ausgangswert)
– Kovarianzanalyse («ANCOVA», Ausgangswert als Kovariable mitführen)
– Keine Signifikanztests gegen den Ausgangswert
Beim Interpretieren von Studienresultaten
– Bewusstsein, dass RZM da ist, wenn nicht ausdrücklich ausgeschlossen
– Nur Aussagen akzeptieren mit dem Zusatz «gegenüber der Vergleichsgruppe/Plazebo»

Leerphase verglichen. Des Weiteren kann man einen Teil der durch die RZM verursachten Merkmalsverbesserung dadurch «wegrechnen», dass man den Ausgangswert als Kovariable in die Analyse einschliesst. Die wichtigste Massnahme ist aber immer die Mitführung einer Vergleichsgruppe mit randomisierter Zuteilung und eine Disziplinierung der Aussagen, die sich nicht auf die Vorwerte, sondern auf die Vergleichsgruppe beziehen sollen. Also statt: «Medikament A senkte den Blutdruck um 12 mm Hg gegenüber dem Ausgangswert» besser: «Medikament A senkte den Blutdruck um 3 mm Hg stärker als Plazebo» [2]. Allerdings ist auch diese Ausdrucksweise salopp verkürzt, da sie impliziert, dass Plazebo den Blutdruck gesenkt hätte. Die beobachtete Veränderung im Plazebo-Arm einer Studie setzt sich aber zusammen aus der eigentlichen Plazebo-*Wirkung* und statistischen Faktoren. Also war der «Übeltäter» nicht (nur) die Plazebo-Tablette ohne Wirkstoff, sondern der Studienleiter, der für seine Untersuchung Personen mit einem bestimmten Krankheitsschweregrad selektioniert hat, in Zusammenarbeit mit den Parzen (Schicksalsgöttinnen), die für die notwendigen Schwankungen des Messparameters gesorgt hatten.

Wir Leser von wissenschaftlichen Arbeiten haben natürlich keinen Einfluss auf das Studiendesign, jedoch können wir die gelesenen Arbeiten danach beurteilen, ob die Autoren der RZM die notwendige Bedeutung beigemessen, das Studiendesign entsprechend angepasst und die Schlussfolgerungen mit der notwendigen Zurückhaltung gezogen haben.

Schliesslich müssen wir die RZM von anderen Effekten abgrenzen, die in ihrer Summe ebenfalls Verlaufsparameter beeinflussen können. Eine beobachtete Medikamentenwirkung setzt sich zusammen aus stochastischen Schwankungen, psychologischen (Kontext-) Faktoren und der eigentlichen Substanzwirkung [3]. Der Hawthorne-Effekt [4] beschreibt den Umstand, dass es Versuchspersonen besser geht, wenn sie in eine wissenschaftliche Studie eingeschlossen sind und damit Aufmerksamkeit und Zuwendung erhalten. Auch zu beobachten ist das Phänomen der «sozialen Erwünschtheit der Antworten», das Versuchspersonen dazu bringen kann, ihren Zustand als besser zu beschreiben, als dies tatsächlich der Fall ist. Ebenso die Uminterpretation von Beschwerden und das Umgehen damit (engl. «coping») ist ein Faktor, der Studienresultate beeinflussen kann. Schliesslich kann eine fehlende Verblindungsmöglichkeit die Ergebnisse einer Untersuchung stören.

Konklusion

Die RZM ist ein statistisches Phänomen, das in allen wissenschaftlichen Untersuchungen eine relevante Rolle spielt, sofern a) eine Selektion des Zielparameters stattfindet, b) eine Verlaufsmessung desselben erfolgt und c) der Zielparameter spontanen Schwankungen unterworfen ist. Die RZM führt dazu, dass die Wirkstärke von Interventionen überschätzt wird, sofern diese nicht mit Plazebo oder einer anderen wirksamen Therapie verglichen werden.

Korrespondenz:
Dr. med. Markus Gnädinger
Birkenweg 8
CH-9323 Steinach
markus.gnaedinger[at]hin.ch

Literatur

1 Galton F: Regression towards mediocrity in hereditary stature. The Journal of the Anthropological Institute of Great Britain and Ireland (The Journal of the Anthropological Institute of Great Britain and Ireland). 1886;15:246–63.
2 Barnett AG, van der Pols JC, Dobson AJ: Regression to the mean: what it is and how to deal with it. Int J Epidemiol. 2005;34:215–20.
3 Gnädinger M, Fässler M: Placebo in der hausärztlichen Praxis. pharmakritik. 2011;33(4):13–6.
4 Kleist P: Vier Effekte, Phänomene und Paradoxe in der Medizin. SMF. 2006;6:1023–7.