Ich hatte im vorigen Kapitel die theoretischen Voraussetzungen des hierarchischen Modells analysiert und eine Alternative, das zirkuläre Modell, vorgeschlagen. Jetzt wollen wir das Ganze etwas vertiefen und überlegen, welche konkreten Konsequenzen sich aus den beiden Modellen ergeben. Ich halte das hierarchische Modell für untauglich. Daraus mache ich keinen Hehl. Ich werde dann im nächsten Kapitel an ein paar Beispielen zeigen, dass das gegenwärtige Modell schlecht funktioniert und auf Dauer zu teuer und wissenschaftlich unbefriedigend ist.

Die praktische Konsequenz des hierarchischen Modells

Der Vorteil der klassischen Strategie: Das Experiment
Wenn wir der Vorgabe des hierarchischen Modells folgen, dann müssen wir sobald als möglich im Forschungsprozess experimentieren, um den „wahren" Effekt einer Intervention zu finden. Ich setze „wahr" in Anführungszeichen, weil ich der Meinung bin, dass diese „Wahrheit" in diesem Kontext eine Fiktion ist. (Das heisst nicht, dass es keine Wahrheit gibt. Schon der Hl. Augustinus hat in einem simplen Argument gezeigt, dass es Wahrheit als Leitidee geben muss: Selbst derjenige, der sagt, es gäbe keine Wahrheit, beansprucht für diese seine Aussage Wahrheit. Also muss es Wahrheit als Grenzidee geben.) Im medizinischen Kontext ist es allerdings eine Fiktion zu glauben, es gäbe eine Wahrheit, die für alle Menschen, in allen Kontexten und unter allen Umständen, in allen Kulturen und zu allen Zeiten und angewandt von allen Therapeuten gleich wirksam ist. Die Standardmeinung geht jedenfalls davon aus oder zumindest wird dies suggeriert, wenn man Aussagen liest wie: „xyz Therapie verbessert die Rückfallquote bei chronisch Depressiven um 38%". Im hierarchischen Modell werden, wie früher kurz beschrieben, nach Möglichkeit experimentell erzeugte Daten verwendet, um solche Aussagen zu erzeugen, weil diese klarere Schlussfolgerungen zulassen.

Warum? Stellen Sie sich vor, Sie hätten zwei Therapien zur Depressionsbehandlung: "Muckelfucktherapie" und Psychopharmaka. Stellen Sie sich vor, Sie hätten zwei grosse Gruppen von Patienten, solche, die sich für Muckelfucktherapie entscheiden, und solche, die lieber Psychopharmaka nehmen. Nun stellen Sie nach einer gewissen Beobachtungszeit fest, dass es den Patienten, die Muckelfucktherapie genommen haben, besser geht als den anderen. Können wir den Unterschied auf die Therapie zurückführen? Nicht notwendigerweise. Denn es könnte ja sein, dass z.B. alle oder viele Patienten, die sich für Muckelfucktherapie entscheiden, eine bestimmte noch nicht bekannte genetische Veranlagung haben, die dazu führt, dass sie Omega-3 Fettsäuren besser verstoffwechseln können, und dass Depression u.a. auch darauf zurückzuführen ist, dass Menschen zuwenig solcher Fettsäuren haben. Wir hätten also in unserer Muckelfuckgruppe implizit solche Menschen, die vielleicht etwas leichter von selber wieder aus ihrer Depression finden und würden eine spontane Besserung fälschlicherweise der Therapie zuschreiben. Oder

Patienten der Muckelfuckgruppe könnten etwas gebildeter sein. Nun wissen wir aber, dass sich besser gebildete Menschen rascher eigene Ressourcen zur Besserung erschliessen können. Also würden wir einen Effekt der sozialen Unterschiede übersehen, wenn wir davon ausgingen, dass die Unterschiede zwischen den Gruppen auf die Therapie zurückzuführen seien.

Es gibt eine Unzahl von möglichen Einflussfaktoren auf Krankheiten. Solche die wir kennen, wie etwa einige genetische Faktoren des Stoffwechsels, Bildung, sozialer Status, Rauchen oder Alkoholkonsum, und viele, die wir nicht kennen. Wer weiss, vielleicht stellt sich irgendwann heraus, dass eine Geburt im Winterhalbjahr ein Risikofaktor in Zusammenhang mit einer bestimmten genetischen Konstellation für irgendeine Krankheit ist?

Randomisierung
Um solche bekannten und auch unbekannten Faktoren in den Griff zu bekommen, wenden Forscher gerne einen Trick an: sie weisen die Patienten auf die Gruppen per Zufall zu, also technisch mit einem Computerprogramm. Dadurch werden alle möglichen Einflussfaktoren so auf beide Gruppen verteilt, dass sie überall einen gleich grossen oder kleinen Einfluss ausüben. Wenn man dann eine Intervention einführt, die nur eine Gruppe erhält, und wenn man sorgfältig misst, dann kann man mit einiger Sicherheit davon ausgehen, dass Unterschiede zwischen den Gruppen mit der Intervention zu tun haben und nicht mit Unterschieden, die schon vorher oder implizit da waren. Diese Theorie greift auf jeden Fall dann, wenn die Studien gross genug sind, also so ca. ab 300 Patienten, und wenn der Zufall bei seiner Ausübung nicht gestört wird, wenn man also unbeschränkt zuteilen würde. Letzteres wird selten gemacht. Denn wenn man einfach nur würfelt, dann kann es sein, dass die Gruppen ungleich gross werden. Das versucht man zu vermeiden, da statistisch betrachtet immer die kleinste Gruppe bestimmt, wie mächtig der Test ist. Wenn man also einen Unterschied von 50 Personen zwischen zwei Gruppen hat, hat man z.B. in der einen Gruppe 150 und in der anderen Gruppe 200, dann hat man 50 Personen umsonst rekrutiert. Da das Einschliessen von Patienten in Studien teuer ist versucht man solche Unterschiede zu vermeiden und randomisiert in Blöcken. Das heisst man beschränkt den Zufall auf Einheiten von 4 oder 8 oder 10 o.ä., so dass sich die Gruppen maximal um so viele Patienten unterscheiden können. Allerdings ist dann aber auch die Zufallszuteilung in ihrer Mächtigkeit beschnitten. Aus diesem Grund funktioniert Randomisation wirklich gut erst ab ca. 150 Patienten pro Gruppe. Es wurden zwar Alternativen vorgeschlagen, die sog. Minimierungsstrategie, bei der Computerprogramme Patienten durch Berechnung der Unterschiede zwischen Gruppen verteilen, aber diese haben sich leider nicht durchgesetzt, weil sie etwas komplizierter sind.

Randomisation führt also dazu, zumindest theoretisch und praktisch in grossen Studien, dass Ausgangswerte in beiden Gruppen gleich verteilt sind. Reicht aber Randomisation

schon aus? Meistens nicht.

Homogenisieren
Meistens führen Forscher noch eine Reihe anderer Methoden ein, um ihre Studien abzusichern. Vor allem versuchen sie, homogene Gruppen zu erzeugen. Warum? Weil sie dann mit kleineren Patientenzahlen Effekte zeigen können. Erinnern wir uns: Patienten in Studien einzuschliessen ist teuer. Manche schätzen, ein Patient kostet in einer längeren Studie bis zu $ 28.000 (das sind Kosten für den Arzt, der eine Prämie kriegt, für wissenschaftliches und ärztliches Personal, das Daten erhebt, auswertet und überwacht, etc.). Man versucht also normalerweise mit möglichst wenig Patienten auszukommen. Das ist schon ethisch notwendig, denn schliesslich ist jedes Experiment immer auch mit Belastungen, möglichen Nachteilen oder Nebenwirkungen verbunden, und Ethikkommissionen achten darauf, dass nicht unnötigerweise experimentiert wird. Wie kann man aber das feine Signal einer Intervention vom Rauschen der Kontrollgruppe trennen? Man arbeitet mit möglichst homogenen Gruppen. Das wird bewerkstelligt, indem man Kriterien formuliert, unter denen man davon ausgeht, dass eine Therapie am besten funktioniert. Ausschlusskriterien sagen, welche Patienten nicht in der Studie behandelt wurden. Häufig finden sich unter diesen Kriterien abgesehen von Standardkriterien wie der Altersbegrenzung, der schwangeren und stillenden Frauen (weil man nicht weiss, ob nicht möglicherweise eine Gefährung eintreten kann) oder des Sprachverständnisses solche, bei denen Patienten mit bestimmen Schweregraden einer Diagnose ausgeschlossen werden – z.B. besonders schwer Depressive, oder leicht Depressive – oder Patienten mit mehreren Diagnosen – z.B. mit Depression und Angst, Abhängigkeitsstörung oder Persönlichkeitsstörung. Das hat zur Folge, dass es meistens leichter ist, in solchen experimentellen Studien Effekte zu erzeugen, die grösser sind als diejenigen in der Kontrollbedingung – oder gleich gut, je nachdem, welche Kontrollbedingung gewählt wird und was man zeigen will.

Andere formale und inhaltliche Voraussetzungen des Experiments
Experimente kann und darf man an Menschen nur durchführen, wenn es gute Gründe dafür gibt. Eine der hauptsächlichen Vorbedingungen ist, dass man nicht genau weiss, was wirklich gut funktioniert, dass also unsere Erkenntnis in der Schwebe ist („equipoise“). Das ist immer dann der Fall, wenn man neue Interventionen testet, von denen keiner weiss, wie gut sie sind. Eine Konsequenz dieser Situation ist, dass keiner, Behandler und Patienten, eine wirkliche Präferenz hat oder haben sollte, die sie zu einer bestimmten Behandlung hingezogen sein lässt. Experimente darf man auch nur dann durchführen, wenn die Patienten wissen, worauf sie sich einlassen und zustimmen, also bewusst ihr Einverständnis geben. Praktisch sieht dies so aus, dass man Patienten schriftlich und mündlich ausführlich erklärt, wie die Studie aufgebaut ist, was alles passiert, wie oft sie kommen müssen, welche Fragebögen sie wann ausfüllen müssen, welche Vorteile, und welche Nachteile sie zu

erwarten haben, welche Messungen vorgenommen werden, wie die Bedingungen aussehen die getestet werden (z.B. Therapie und Placebo, oder zwei verschiedene Therapien) – und mit welchen Nebenwirkungen zu rechnen ist. Des weiteren kann man solche Studien meistens nur mit einer entsprechenden Logistik aufbauen. Die findet sich aber nur bei grossen Kliniken, in Universitäten oder bei spezialisierten Unternehmen. Schätzungen gehen davon aus, dass nur ungefähr 1-5% aller Patienten in klinischen Studien aus der niedergelassenen Praxis kommen, der Rest wird in Kliniken, also in spezialisierten Behandlungszentren rekrutiert.

Das führt dazu, dass nur bestimmte Patienten in Studien eingeschlossen werden: solche, denen es egal ist, wie sie behandelt werden und die voll und ganz der Klinik, dem Studienzentrum oder dem Arzt vertrauen und solche, die mit Erkrankungen, die in der niedergelassenen Praxis nicht mehr behandelbar sind, in der Klinik landen.

Der Nachteil der klassischen Strategie: mangelnde Generalisierbarkeit
Daran erkennt man den hauptsächlichen Nachteil dieser experimentellen Strategie: die Ergebnisse sind streng genommen nur auf eine ganz kleine Zahl aller Patienten anwendbar. Bei 95% aller Patienten wissen wir nicht, ob die gefundenen Ergebnisse überhaupt anwendbar sind. Dies ist das Problem der Generalisierbarkeit oder der sog. „externen Validität“. Das Schlimme daran ist folgendes: Wir wissen nicht genau, wie interne Validität, also die methodischen Charakteristika einer Studie, und externe Validität, also die Generalisierbarkeit auf andere Patienten miteinander zusammenhängen und können daher nicht durch mathematische Modelle oder Überlegungen dieses Manko wettmachen. Wir wissen nur eines: je höher die interne Validität ist, umso grösser ist die Wahrscheinlichkeit, dass die externe Validität sinkt. Denn mit jedem ausgeschlossenen Patienten, mit jedem Ausschlusskriterium, mit jedem Patienten der keine Lust hat, durch Zufall einer Behandlung zugeteilt zu werden; mit jedem Patienten, der nicht in einem spezialisierten Studienzentrum behandelt wird sinkt die Generalisierbarkeit. Dies ist weniger ein Problem für extrem dicht beforschte Gebiete, wie etwa die akute Onkologie. Da wissen wir meistens sehr gut, was wie funktioniert, denn hier werden die Patienten tatsächlich dort rekrutiert, wo sie auch behandelt werden. Es ist aber ein grosses Problem für alle eher vagen Erkrankungen oder für Erkrankungen, die oft mit verschiedenen anderen Diagnosen einhergehen. Und das sind die allermeisten anderen Erkrankungen.

Ich will dies an einem Beispiel verdeutlichen: Wir haben eine Fülle von psychopharmakologischen Depressionstherapien. Sie sind alle amtlich zugelassen, haben also irgendwann einmal mindestens eine, in der Regel mehrere, Studien hinter sich gebracht, die zeigen, dass sie einer Scheintherapie, in diesem Falle Placebo, überlegen waren. Für fast alle gibt es auch eine Fülle von Studien, die zeigten, dass sie nicht besser als Placebo waren, genauer gesagt in mehr als der Hälfte der Fälle war das der Fall, aber

grosso modo funktionieren sie. Die Effekte sind nicht überragend gross, aber alles zusammen, Placebo-Effekt und pharmakologischer Effekt ist in diesen Studien gross genug, so dass man den Eindruck gewinnt, die Medikamente funktionieren (die Frage nach dem Placebo-Effekt behandeln wir später). Nun wurden diese Daten alle in gezielten Experimenten gewonnen: mit Patienten, die nur Depression hatten, nichts anderes, und zwar nicht zu stark und nicht zu wenig depressiv, die keine Alkoholabhängigkeit hatten, wo die Depression nicht als Folgeerscheinung anderer Erkrankungen auftrat, die keine zusätzliche Angststörung hatten etc.

In der Praxis haben aber die meisten Depressiven noch viele andere Probleme. Deswegen hat man eine riesige Studie angestrebt, die die Effekte von Depressionstherapie untersucht hat, so wie sie in der Praxis stattfindet, die sog. STAR*D-Studie: In einem ausgeklügelten Eskalationsprogramm konnten Psychiater von einer Medikation zur nächsten wechseln, wenn die erste nicht funktioniert hat, auch Psychotherapie verordnen, bis am Schluss ganz neue, starke und auch nebenwirkungsträchtige Medikamente zum Einsatz kamen, ganz so wie auch in der Praxis. Das Ergebnis war ernüchternd: weniger als 50% der Patienten werden dauerhaft (in diesem Falle mindestens ein Jahr) frei von ihrer Depression. Eine kritische Analyse zeigt sogar, dass die Daten geschönt wurden und insgesamt weniger als 38% von dieser pharmakologischen Therapie profitieren. Dieses Beispiel zeigt: was man aus randomisierten, klinischen Experimenten an Erkenntnissen gewinnt, ist nicht notwendigerweise auf die Praxis anwendbar – eben weil die Generalisierbarkeit der Ergebnisse durch das Experimentieren selbst eingeschränkt wird.

Wir müssen also immer zwischen Scylla und Charybdis durchsegeln: auf der einen Seite wollen wir gültige Ergebnisse, auf der anderen Seite wollen wir Ergebnisse, die anwendbar sind. Kann man das nicht in einer richtig guten Studie gemeinsam klären? Jein. Man könnte, in sog. „Megatrials – Riesenstudien“ meinethalben 100.000 Leute zufällig auf zwei Bedingungen aufteilen und behandeln, keine Ausschlusskriterien ausser der Diagnose. Dann hätte man maximal generalisierbare, experimentelle Daten. Das Problem: solche Studien sind extrem teuer und etwa in Europa kaum durchführbar. Daher weichen Proponenten solcher Studien nach Russland, China oder anderswo aus. Können wir dann solche Ergebnisse in Europa verwenden? Keiner weiss es. Ausserdem könnte es sein, dass eine wertvolle Behandlungsmethode nur bei einer bestimmten Gruppe von Patienten funktioniert. Solche differenzierten Effekte werden in Riesenstudien übersehen. Daher kann man keine eierlegende Wollmilchsau erfinden, die zugleich gültige und generalisierbare Ergebnisse liefert. Vielmehr muss man auf eine Strategie ausweichen, die diese Daten in unterschiedlichen Studien erzeugt und dann zusammenführt. Dies ist es genau, was das zirkuläre Modell vorschlägt.

← Zurück zu Kapitel 2

Weiter zu Kapitel 4 →

- Literatur:
  Aickin, M. (1983). Some large trial properties of minimum likelihood allocation. Journal of Statistical Planning and Inference, 8, 11-20.
- Aickin, M. (2001). Randomization, balance, and the validity and efficiency of design-adaptive allocation methods. Journal of Statistical Planning and Inference 94, 97-119.
- Aickin, M. (2002). Beyond randomization. Journal of Alternative and Complementary Medicine, 8, 765-772.
- Fava, G. A., Tomba, E., & Grandi, S. (2007). The road to recovery from depression – don´t drive today with yesterday´s map. Psychotherapy and Psychosomatics, 76, 260-265.
- Khan, A., Khan, S., & Brown, W. A. (2002). Are placebo controls necessary to test new antidepressants and anxiolytics? International Journal of Neuropsychopharmacology, 5, 193-197.
- Pigott, H. E., Leventhal, A. M., Alter, G. S., & Boren, J. J. (2010). Efficacy and effectiveness of antidepressants: current status of research. Psychotherapy and Psychosomatics, 79, 267-279.
- Rush, J. A., Trivedi, M. H., Wisniewski, S. R., Nierenberg, A. A., Stewart, J. W., Warden, D., et al. (2006). Acute and longer-term outcomes in depressed outpatients requiring one or several treatment steps: A STAR*D report American Journal of Psychiatry, 163, 1905-1917
- Stewart, D. J., Whitney, S. N., & Kurzrock, R. (2010). Equipoise lost: ethics, costs, and the regulation of cancer clinical research. Journal of Clinical Oncology, 28, 2925-2935.
- Walach, H., Falkenberg, T., Fonnebo, V., Lewith, G., & Jonas, W. (2006). Circular instead of hierarchical – Methodological principles for the evaluation of complex interventions. BMC Medical Research Methodology, 6(29).