Prof. Dr. Franz Porzsolt, Leiter Versorgungsforschung (Klinik f. Allgemein-/Viszeralchirurgie, Uni-Klinikum Ulm)

„Externe Evidenz ist genauso wichtig wie interne“

Der Onkologe und Internist Prof. Dr. Franz Porzsolt wird bei den Einen als Nestbeschmutzer angesehen, bei den Anderen als kritischer Geist, weil er kein Blatt vor den Mund nimmt, wenn die von ihm so hochgeschätzte Evidenzbasierte Medizin seiner Meinung nach unzulässig verkürzt wird auf Randomisation. Man muss nicht immer seiner Meinung sein, doch hat er sicher recht, wenn er sagt, „dass jede Wissenschaftsrichtung interne Kritik braucht, um ernst genommen zu werden“. Das gilt umso mehr, wenn Vertreter von Institutionen eine Generalkritik an Registern zu üben versuchen, ohne dabei so manche Unzulänglichkeiten der RCT zu erwähnen (s. S. 16). Das hingegen tut Porzsolt in diesem Interview, um in „Monitor Versorgungsforschung“ eine offene, sicher auch kontroverse Fachdiskussion zu eröffnen, mit dem erklärten Ziel, zu einem besseren Verständnis und Umgang mit dem eigentlich „recht umfangreichen Werkzeugkasten der EBM“ zu kommen.

>> Herr Prof. Porzolt, wenn Sie einmal RCT und Pragmatic Trials ...

... Stopp. Das ist so, als wenn ein Handwerker versuchen würde, wechselseitig einen Schraubenzieher und eine Kneifzange für den gleichen Arbeitsgang zu verwenden. Kein einigermaßen ausgebildeter Handwerker wird sagen: Reiche mir mal einen Schraubenzieher oder eine Beißzange, ich komme damit schon irgendwie zurecht.

Was sind die Unterschiedlichkeiten beider Instrumente?

Beider? Die Grundlage der Evidence-based Medicine ist die klinische Epidemiologie, die ein ganzes Bündel an naturwissenschaftlichem Regelwerk beinhaltet, oder – um im Bild zu bleiben – einen recht umfangreichen Werkzeugkasten vorgibt, dessen Instrumente man nutzen muss, wenn man eine Studie plant und durchführt. Dazu sollte man jedoch die Vor- und Nachteile sowie Limitationen der jeweiligen Instrumente sehr genau kennen. Das große Problem der letzten 25 Jahren – denn so alt ist Evidence-based Medicine inzwischen – ist es, dass viele Wissenschaftler den Baukasten an Regelmaßnahmen im Prinzip auf die Randomisation verkürzen. Damit sage ich rein gar nichts gegen die Randomisierung als solche, sondern nur gegen die damit fast zwanghaft einhergehende Standardmeinung, dass alles gut ist, was eine randomisierte Studie ist, und im Umkehrschluss, alles was nicht randomisiert ist, demnach schlecht sein muss.

Ist das nicht auch so etwas wie ein Implicit Bias?

Der kommt doch noch oben drauf. Dieser Bias entsteht alleine schon durch die Art und Weise, wie wir Dinge wahrnehmen. Der Implicit Bias entstammt eigentlich der US-Soziologie, die in Studien oft das Problem hatte, dass schlechte Eigenschaften fast automatisch Black People, eher gute hingegen White People zugeschrieben werden, was natürlich völliger Blödsinn ist. Diesen Implicit Bias gibt es nun aber auch in vielen anderen Bereichen der Gesundheitsforschung, weil zum Beispiel ein Endokrinologe – wenn ein Patient über Beschwerden berichtet – diese fast automatisch auf Diabetes zurückführt, während ein Rheumatologe die gleichen berichteten Symptome wahrscheinlich mit rheumatischen Problemen erklären wird. Gleiches beobachtet man übrigens auch in Tumorkonzilen, die ich jahrelang durchgeführt habe: Wenn es die Möglichkeit gab, einen Tumor zu bestrahlen, war der Strahlentherapeut immer dafür, währenddessen der Chirurg den Patienten lieber gleich operieren wollte und der Internist eher eine Chemotherapie präferierte. Das ist aber wiederum ganz verständlich, weil jeder Arzt nun einmal die Vor- und Nachteile einer Therapie aus dem eigenen Kompetenzbereich einfach besser als jene kennt, die aus anderen stammen.

Und was denken Sie als Onkologe und Internist?

Ich habe in solchen Tumorkonzilen vor so manchen Chemotherapien gewarnt, einfach deshalb, weil sie oft mehr schaden als nützen.

Damit gilt man in der jeweiligen Fachcommunity doch schnell als Nestbeschmutzer, oder?

Sicher. Aber nur, weil viele nicht realisiert haben, dass jedes Fach und jede Wissenschaftsrichtung interne Kritik braucht, um überleben zu können und auch, um ernst genommen zu werden. Ich habe nie versucht, meine eigene Disziplin kaputt zu machen, sondern hatte immer genau das Gegenteil im Sinn, weil man Internisten – wie Vertretern aller anderen Fachrichtungen – doch nur dann vertrauen kann, wenn sie sehr sorgfältig abwägen, welche – auch fachfremde – Therapien den bestmöglichen Outcome bringen.

„Nahezu niemand kennt die Ausgangsrisiken der zu behandelnden oder in eine Studie einzuschließenden Patienten. Übrigens auch nicht in Disease-Management-Programmen!“

Was zu beweisen wäre.

Genau da liegt der Hase im Pfeffer. Denn: Ist denn überhaupt klar, welche Therapien den bestmöglichen Outcome bringen? Auf den ersten Blick ist das vielleicht so, doch dann stößt man ziemlich schnell auf das Kardinalproblem jedweder therapeutisch-medizinischen Realität und damit auch jeder klinischen Studie: Nahezu niemand kennt die Ausgangsrisiken der zu behandelnden oder in eine Studie einzuschließenden Patienten. Übrigens auch nicht in Disease-Management-Programmen.

Darum ist es ja so schwer zu evaluieren, ob sie nun funktionieren oder nicht.

Aber doch nur deshalb, weil vorher keiner auf die Idee gekommen ist, die Ausgangsrisiken zu erfassen. Daher nützen die ganzen zu DMP publizierten Ergebnisse eigentlich wenig, weil man doch schon von vorneherein voraussagen kann, dass der Patient mit einem hohen Risiko wohl ein schlechteres Ergebnis haben wird, als der mit einem niedrigeren. Es gilt: In solchen Studien dürfen nur risikogleiche Patienten miteinander verglichen werden. Das ist das A und O, eigentlich ein recht einfacher Gedanke, der leider nur zu oft nicht gedacht wird.

Das Problem der Ausgangslast stellt sich doch auch bei Mindestmengen.

Natürlich. Das Problem der Mindestmengen kann man überhaupt

nur dann angehen, wenn man in der Lage ist, vergleichbare Risiken herzustellen. Wenn ein kleines Krankenhaus einen Patienten mit extrem hohem Risiko eingeliefert bekommt, erkennt man schnell, ob ein guter Arzt die Aufnahme macht, frei nach Sauerbruch: „Der gute Arzt unterscheidet sich vom Stümper dadurch, dass der gute den schwierigen Patienten erkennt." Dagegen wird der unerfahrene Arzt, den Patienten mit schlechter Prognose genauso behandeln wie den mit einer guten Prognose und wird alleine schon dadurch ein schlechteres Ergebnis erzeugen. Das führt zum nächsten Bias: Ein guter Arzt in einem kleinen Krankenhaus wird im Sinne des Patienten die hohen Risiken aussortieren und an ein besseres, oft größeres weiterreichen. Es gibt schon eine ganze Menge an Faktoren, die an sich ganz logisch sind, aber an die in der bisherigen Qualitätsdebatte kaum jemand denkt.

Und warum denkt man daran nicht?

Viele Forscher und vor allem Institutionen wurden verführt durch die scheinbare Sicherheit, die Randomized Controlled Trials, kurz RCT, versprechen.

Wie das?

Das ist ganz einfach: Im RCT muss man nicht mehr groß über Risiken nachdenken, weil alleine schon durch die Randomisierung gewährleistet wird, dass die Risiken in der Interventions- und Kontrollgruppe gleich verteilt werden.

Aber damit doch auch gleich ausgeblendet werden.

Genau das ist das Problem, denn das hat den traurigen Nebeneffekt, dass eigentlich niemand mehr über die Risiken als solche nachdenkt.

Ein Beispiel?

Stellen wir uns einmal vor, dass wir zwei Studien durchführen, eine beispielsweise in Bonn und eine in Rosenheim, nur um zwei annähernd gleichgroße Städte zu nehmen. Beide Studien sind randomisiert und beide behandeln im Prinzip die gleiche Erkrankung. Nun kann es aber sein, dass in der Bonner Klinik vornehmlich Schwerkranke behandelt worden sind, und in Rosenheim vorwiegend leichtere Fälle, weil die schweren – durch den vorher angenommenen guten Arzt – sowieso schon alle nach München geschickt worden sind. Damit wurde zwar ausreichend randomisiert, doch ist in der Bonner Kohorte das durchschnittliche Risiko erheblich höher als in der Rosenheimer. Das heißt nichts anderes, als dass es verdammt schwierig bis nahezu unmöglich ist, die Ergebnisse beider Studien miteinander zu vergleichen. Und nur deshalb, weil der verlässliche Parameter des vorhandenen, und an sich ganz leicht zu erhebenden Ausgangs-Durchschnittsrisikos fehlt.

Und wenn nun auch die Bonner Klinik – den guten Arzt wieder vorausgesetzt – die schwierigeren Fälle nach Köln schickt?

Wenn auch die leichten und die schweren woanders behandelt werden, wird es sich um relativ homogene Gruppen von mittleren Risiken handeln. Damit wären die Gruppen wieder vergleichbar. Doch das wahre Problem liegt wo anders verortet: Durch die Randomisierung macht sich der Studiendurchführende über das Ausgangsrisiko gar keine Gedanken mehr, weil er immer nur die interne Evidenz – also die Evidenz, die innerhalb der Studie existiert – vor Augen hat. Damit wird ihm durch die Randomisierung vorgegaukelt, dass in beiden Ausgangsgruppen eine ähnliche oder gleiche Risikoverteilung

Bild aus Video: https://drive.google.com/file/d/0B9MPdcimmmmpQkQ0cjdKNDk0R0U/view



vorhanden ist, was aber nicht stimmt.

Nun werden zumindest in der frühen Nutzenbewertung keine anderen als RCT-Studien zugelassen, weil sie nun einmal die höchste interne Evidenzklasse haben. Punkt.

Damit hat das IQWiG ja auch recht. Womit das Institut jedoch nicht recht hat, ist die Annahme, dass die interne Evidenz wichtiger als die externe ist. Ich bin sogar der festen Überzeugung, dass die externe Evidenz mindestens genauso wichtig, wenn nicht vielleicht sogar wichtiger ist.

Doch ist diese Meinung nicht bewiesen.

Das ist das Problem der Fehlerabschätzung.

Will heißen?

Nehmen wir wieder ein einfaches Beispiel: Wenn bei einem Auto auf einmal der Motor nicht mehr funktioniert, oder aber ein Reifen platzt, hat der Autofahrer wohl ein Problem auf seiner Fahrt. Dabei ist es ihm völlig egal, ob dafür die interne Evidenz (eben der Motor) schuld ist oder die externe Evidenz (der Reifen) nicht in Ordnung ist. Wenn er die Fahrt fortsetzen möchte, muss er beide in funktionstüchtigem Zustand haben.

Übertragen auf Studien heißt das?

Im Prinzip ist das ein Totschlagargument für fast jede Studie: Wenn die interne Evidenz stimmt, kannst man den gefundenen Unterschied glauben oder nicht glauben. Doch wenn die dazu gehörige externe Evidenz nicht stimmt, kann man das gefundene Ergebnis auf keinen Fall außerhalb der Studienpopulation anwenden. Das heißt: Der gefundene Effekt mag zwar hochinteressant für die meist recht enge Studienpopulation sein, doch gilt er nicht für den großen Rest der Welt.

Das scheint ein großes Problem zu sein, das recht selten thema-

tisiert wird.

Warum wohl? Damit wäre doch die Gläubigkeit an RCT dahin, oder?

RCT charakterisieren sich aber nun auch durch ihre Ein- und Ausschlusskriterien.

Stimmt, davon gibt es jede Menge. Im „New England Journal" veröffentlichte Studien kommen gut und gern auf 10 bis 20 Ein- und Ausschlusskriterien.

„Die Ausschlusskriterien richten sich nach den Risiken, und stehen eben nicht zu 100 Prozent mit der zu testenden Innovation oder Therapie in Zusammenhang."

Beginnen wir bei den Ausschlusskriterien, also all dem, was nicht in eine Studie darf.

Ausschlusskriterien haben die originäre Aufgabe, eine möglichst eng begrenzte Population zu erzeugen, die gerade groß genug ist, dass sie für die gewünschte Messung noch geeignet ist. Wenn man streng logisch denkt, muss man zu dem Ergebnis kommen, dass damit all das, was man in dieser Studie an Erkenntnissen findet, nur anwendbar auf die betrachtete Population ist – das ist aber nichts neues, das haben wir alle beim leider inzwischen verstorbenen Vater der EBM, David Sackett, gelernt.

Wenn ich nun zum Beispiel eine Studie im Bereich des diabetischen Fußsyndroms aufsetzen möchte, wäre es doch sinnvoll, nur jene Ein- und Ausschlusskriterien zu nehmen, die durch die Fachgesellschaften konsentiert worden sind.

Wäre es. Doch kann die Ein- und Ausschlusskriterien jeder Studiendurchführende für seine Studie selbst festlegen, was zum Teil aber auch ganz sinnvoll ist, weil beispielsweise Komorbiditäten für Patienten brandgefährlich sein können – darum gehören diese ausgeschlossen, auch wenn dadurch das Ergebnis verzerrt wird. Das heißt aber auch, dass sich die Ausschlusskriterien nach den Risiken richten, und eben nicht zu 100 Prozent mit der zu testenden Innovation oder Therapie in Zusammenhang stehen.

Es wird aber doch sicherlich Ein- und Ausschlusskriterien geben, die für eine bestimmte Population, die mit einem bestimmten Medikament oder einer Therapieform behandelt werden sollen, am sinnvollsten sind.

In gewisser Weise kann man sich durchaus vorstellen, dass die Ein- und Ausschlusskriterien standardisierbar sind. Aber leider ist es so nicht. Es gibt tausend Gründe, Ausschlusskriterien zu formulieren, manche vielleicht aus der Angst des hinter einer Studie stehenden Unternehmens heraus, dass das zu testende Arzneimittel überhaupt nicht mehr wirkt; oder dass Intensivpatienten ganz ausgeschlossen werden, weil der Verdacht aufkommen könnte, dass das zu testende Arzneimittel die Patienten auf die Intensivstation treiben würde. Das ist natürlich alles Quatsch, ist aber die Realität, die nur dadurch existieren kann, dass weder Ein- noch Ausschlusskriterien begründet werden müssen.

Sind denn wenigstens Interventions- und Kontrollgruppen gleich?

Sie sind leider nur annähernd gleich, wie bei allen Studien die sogenannte „Table No. 1" verrät. Ich habe letztens eine Doktorarbeit ausgegraben, die ich 2015 betreut habe, und die eben jene „Table No. 1" untersucht hat.

Klingt ziemlich banal.

Ist es aber nicht. Es gibt in jeder randomisierten Studie diese Tabelle Nr. 1, in der beispielsweise steht: Alter, Gruppe a = 62 Jahre im Durchschnitt. Gruppe b = 62,5. Anteil Männer zu Frauen 1:1 in der Gruppe a, in der Gruppe b 1:1,5. Vorerkrankung Herzinfarkt = 22 % in a, 23 % in b.

Das ist doch auch gut so, denn diese Tabelle Nr. 1 beschreibt nun einmal die Vergleichbarkeit der Ausgangsgruppen.

Zumindest ist das so in 98 und 99 Prozent aller randomisierten und publizierten Studien, der Rest vergisst das einfach. Doch nun hat mein ehemaliger Doktorant ein Experiment durchgeführt, in dem er eine Expertenbefragung zu eben jener Tabelle Nr. 1 durchgeführt hat. Dabei hat er nichts anderes getan, als die prinzipielle Frage zu stellen: Welche Kriterien sollen in Tabelle Nr. 1 erscheinen und welche nicht?

Nun, die Ein- und Ausschlüsse.

Erst in zweiter Linie. Zuerst einmal muss der Endpunkt der Studie festgelegt werden, wobei eine randomisierte Studie immer nur einen einzigen Endpunkt hat – eben den Primary Endpoint. Natürlich kann man auch weitere – Secondary Endpoints – mit untersuchen und sie auch beschreiben, aber die kann man nicht evident testen. Das kann eine randomisierte Studie eben nur mit dem Primary Endpoint, weil sie genau daraufhin gepowert ist. Diesen Primary Endpoint muss man unbedingt zu erst festlegen.

Sagen wir einmal: Tod. Das kann man sehr leicht merken.

Ok, nehmen wir den finalen Endpunkt Tod. Jetzt erst weiß der Studiendurchführende, welche Studienpopulation er braucht und muss dann in Tabelle Nr. 1 alle jene Faktoren inkludieren, die diesen Endpunkt beeinflussen können.

Damit fällt beispielsweise banaler Schnupfen heraus, weil dieser nie tödlich endet. Aber eine maligne Vorerkrankung, wenn die überhaupt zulässig ist, muss rein.

Generell gilt: Wenn man zum Beispiel Herzinfarkte untersuchen will, muss alles rein, was dazu beitragen könnte, diesen Endpunkt zu beeinflussen. An der Stelle frage ich meine Studenten oft: Gehört da auch rein, ob der Patient katholisch oder evangelisch ist?

Die werden sagen: Nein.

Stimmt. Es kann sich keiner vorstellen, dass das Glaubensbekenntnis oder die Religionszugehörigkeit rein muss. Ein Student hat einmal eingeworfen, wenn es um Trichinen geht, schon. Das macht doch Sinn, oder?

Das will aber doch nur sagen, dass in die Tabelle 1 alles, aber auch alles aufgezählt werden muss, was den Endpunkt 1 durch einen wie auch immer gearteten Risikofaktor beeinflussen kann.

Richtig, das wird aber oft vergessen. Denn die so simpel erscheinende Tabelle Nummer 1 reflektiert nichts anderes als die Gleichverteilung der Risiken und diese ist bei jeder Studie mit das Ausschlaggebenste überhaupt.

Und dahinter kommt dann noch das kleine „p", das die statistische Unterschiedlichkeit angibt.

Wenn es denn so wäre, wäre das gut. Nur kenne ich viele Studien,

DocMorris

Medikamente allein sind nicht genug.

GEMEINSAM ZU MEHR THERAPIEERFOLG

indikationsbezogene
Patienten-Begleitprogramme

Zugang zum Patienten



Stärkung der Compliance



Verbesserung der Lebensqualität

WERDEN SIE PARTNER!

E-Mail an: patientenprogramme@docmorris.de

in denen das „p“ in keinem Fall statistisch unterschiedlich war. Dabei kann man schnell eine Art Abakus-Methode anwenden und sich bei jeder Zeile von Tabelle 1 notieren, auf welcher Seite das Risiko ein bisschen höher oder niedriger ist, wobei es meist nie ganz identisch ist.

Als Beispiel nehmen wir einmal ein Durchschnittsalter 60,0 in Gruppe a zu 60,1 in Gruppe b an.

Da würde doch jeder sagen: Das ist doch das gleiche „p“ und damit ist es niemals signifikant, oder?

Stimmt.

Aber nur dann, wenn man nur den einen Wert in einer einzigen Zeile vergleicht. Wer sich aber einmal die Mühe macht, alle Zeilen in Tabelle 1 zu vergleichen, wird feststellen, dass es Studien gibt mit – sagen wir – acht Punkten bei Gruppe a und nur einem bei Gruppe b. Das heißt nichts anderes, als dass sich der vernachlässigbare Unterschied in einer Zeile akkumuliert und in der Zusammenschau aller Risikofaktoren ein deutlich asymmetrisches Verhältnis erzeugt. Daran sieht man, wie wichtig die genaue Betrachtung von Studien ist, und selbst diese scheinbar so einfach daherkommende Tabelle 1 im Endeffekt schon verrät, warum eine Studie ein signifikantes Ergebnis erzeugt hat oder nicht.

„Die scheinbar so einfach daherkommende Tabelle 1 verrät im Endeffekt schon, warum eine Studie ein signifikantes Ergebnis erzeugt hat oder nicht.“

Das führt doch zur Frage, ob RCT – zumindest so, wie sie derzeit oft gemacht sind – Grundlage von EBM sein dürfen?

Jetzt wird es aber politisch. Denn so weit darf man auch wieder nicht gehen. Ich würde diese Frage etwas einschränken wollen und sagen: Sie dürfen nicht länger die alleinige und nahezu unhinterfragte Grundlage sein! Man sollte doch das Kind nicht gleich mit dem Bade ausschütten, denn Randomized Controlled Trials sind nun einmal eine notwendige, aber eben keine hinreichende Bedingung, um daraus Empfehlungen für eine klinische Leitlinie oder auch für die Nutzenbewertung abzuleiten.

Bei der frühen Nutzenbewertung dürften doch, folgt man Ihrer Argumentation, die Ergebnisse nur auf jene Populationen angewandt werden, die auch eingeschlossen waren?

Nicht einmal das. In randomisierte Studien kommen doch nur jene Patienten hinein, die keine starke Präferenz haben.

Präferenz zu was oder wem?

Zu allem und jedem. Nehmen wir eine Studie zur Schlüssellochoperation versus großer Schnitt. Da sagt eine Patientin vielleicht: „Meine Nachbarin hat zu mir gesagt, lass bloß den großen Schnitt nicht machen, das sieht blöd aus, wenn du einen Bikini trägst.“ Dann kommt eben diese Patientin in die Klinik und der Doktor sagt: „Wir machen jetzt eine ganz sorgfältige Operation mit dem großen Schnitt, die ist besser als eine weniger übersichtliche Operation mit dem Schlüsselloch.“

Der Arzt weiß immer ganz genau, was er „verkaufen“ will, oder?

Sicher. Die Ärzte kriegen das schon so hingedreht, dass die Patienten ihnen alles unterschreiben, was sie brauchen. Das bedeutet aber doch nichts anderes, als dass der Patient vom Arzt zu etwas überredet wird, obwohl er vielleicht eine ganz andere Präferenz hat. Dann wird operiert und das ganze geht schief. Woran liegt es dann? An der OP-Methode? Oder vielleicht auch an der Präferenz oder dem Werturteil des Doktors?

Was passiert denn damit im Studiensetting?

Da hat man es meist mit Patientenkohorten zu tun, die fast immer ähnlich verteilt sind: 80 zu 20. 80 Prozent aller Patienten wollen gerne die Schlüssellochoperation, weil diese sozialverträglicher ist als der große Schnitt, während 20 Prozent lieber die große Operation präferieren, weil sie ihnen sicherer erscheint. Nun randomisiert der Studiendurchführende ganz sauber die Patienten in zwei Gruppen und steckt in jede Gruppe brav 80 Prozent mit der Präferenz Schlüssellochoperation und 20 Prozent mit großem Schnitt. Nun wird operiert, zu je 100 Prozent in Gruppe 1 mit Schlüsselloch und zu 100 Prozent in Gruppe 2 mit konventionellem großen Schnitt.

Und was passiert?

Etwas ganz logisches. In der Gruppe 1 mit 80-prozentiger Präferenz zur Schlüsselloch-OP bekommen genau die jene Therapie, die sie von Anfang an wollten. Doch in Gruppe 2 bekommen nur 20 Prozent der Patienten den großen Schnitt, den sie eigentlich präferierten. Das heißt: In der Studie wurden zwar sauber die Therapien randomisiert, aber nicht die Patientenpräferenzen.

Kann sich denn niemand vorstellen, dass dieses psychologische Event einen Einfluss auf das Ergebnis hat?

Es gibt genügend Papers, unter anderem von mir, die das genau belegen. Warum das so ist, steht jedoch bisher in keinem Lehrbuch. Nicht einmal in dem von Sackett.

Könnte das einen Paradigmenwechsel bei der fast monotheistischen Betrachtung von RCT auslösen?

Das glaube ich nicht. Dazu halten die handelnden Organisationen, allen voran das IQWiG und der G-BA, viel zu sehr am klassischen RCT fest. Wie gesagt: Auch mit recht, da RCT der Studienansatz ist, der den geringstmöglichen Beschiss erlaubt.

Was wäre denn Ihr Rat?

Wer ein gutes RCT hat, hat sich nicht mehr, aber auch nicht weniger als die Rechtfertigung erworben, darauf aufsetzend ein Pragmatic Controlled Trial durchführen zu können. Weil das RCT – zumindest wenn die Studie gut aufgesetzt wurde – nichts anderes als die Tatsache bestätigt hat, dass das getestete Prinzip unter Idealbedingungen funktionieren könnte.

Doch dann muss man die zweite, gleichwertige oder vielleicht sogar wichtigere Frage stellen: Funktioniert denn das Prinzip auch unter Alltagsbedingungen?

Nun ist es aber leider so, dass man das bisher gar nicht wissen wollte. Mein Credo lautet daher, dass wir die Ergebnisse, die wir in guten randomisierten Studien erzielt haben, für die Realität nur als Hypothese auffassen dürfen. Denn wenn etwas unter idealisierten Bedingungen funktioniert, heißt das meiner Meinung nach noch lange nicht, dass es auch unter Alltagsbedingungen funktioniert: Aber: Wir haben mit RCT immerhin eine berechtigte Hypothese anzunehmen, dass es unter Alltagsbedingungen ebenso funktionieren könnte.

Kann man denn einen Effekt, der unter Idealbedingungen festgestellt wurde, auch unter Alltagsbedingungen testen? Wo doch die ganzen Confounder und Störfaktoren nicht entfernt werden können?



Genau darum geht es beim sogenannten Pragmatic Controlled Trial, kurz PCT: Während eine randomisierte Studie durch die Definition von Ein- und Ausschlusskriterien charakterisiert ist, besitzt ein Pragmatic Controlled Trial lediglich Einschlusskriterien. Das ist eigentlich auch ganz logisch, denn unter Alltagsbedingungen gibt es eben keine Ausschlüsse.

Werden denn solche Pragmatic Controlled Trials durchgeführt?

Hand aufs Herz: Kein Unternehmen will sich diesen Terz antun, ein PCT nach einem RCT durchzuführen, immer verbunden mit der Gefahr herauszufinden, dass der gefundene kleine Effekt im RCT auf einmal gar nicht mehr nachweisbar ist. Darum habe ich mir einen kleinen Trick überlegt.

Welchen dann?

Einen, der den Unterschied zwischen einem ein- und zweiseitigen Test ausgleicht.

Wie das?

Die meisten publizierten Studien sind einseitig. Das heißt: Ein sinnvoller Vergleich berücksichtigt nur einen der beiden möglichen Vergleiche. Ein Beispiel: Mich interessiert, ob meine neue Therapie nur scheinbar oder tatsächlich besser ist als Placebo. Es ist mir aber egal, ob sie nur scheinbar oder tatsächlich schlechter ist als Placebo. Deshalb prüfe ich nur die Überlegenheit, nicht aber die Unterlegenheit.

Letzteres will niemand wissen, der für eine Studie Geld bezahlt.

Exakt. Bei einem zweiseitigen Test halbiert sich nun einmal gegenüber einem einseitigen Test die Chance, ein signifikantes Ergebnis zu bestätigen, allerdings erhöht sich auch das Risiko, einen Fehler erster Ordnung zu begehen.

Die Mehrzahl der RCT wird demnach einseitig sein.

Nahezu alle sind einseitig. Das ist aber auch ganz vernünftig.

Nun aber haben Sie einen Leserbrief im „New England Journal of Medicine" publiziert, weil es in diesem hoch gerankten Journal ein Paper gab, in dem publiziert wurde, dass bei einem neuen Arzneimittel gegen Diabetes in einem sogenannten Non-Inferiority-Test herausgefunden wurde, dass es gegenüber Placebo nicht unterlegen ist.

Das ist doch aber auch Unsinn, oder? Ein Non-Inferiority-Test macht nur dann Sinn, wenn man zwei wirksame Prinzipien vergleicht. Wenn man beispielsweise zwei annähernd gleich gute Blutdrucksenker hat, die den Blutdruck nachweislich senken, kann man gegenüber Placebo den Nachweis erbringen, dass beide funktionieren. Da aber nun der eine Blutdrucksenker doppelt so teuer war wie der andere, hat man eben einen Non-Inferiority-Test durchgeführt, bei dem definiert wurde, dass der Billigere ein bisschen schlechter als der teurer sein darf, wenn er die Grenze der Non-Inferiorität nicht unterschreitet.

Klingt irgendwie so sinnig wie Leerverkäufe bei Aktien.

So ähnlich, weil all diese Leute anscheinend hoffen, dass ihnen keiner auf die Schliche kommt. Mit diesen Non-Inferiority-Test wollen sie den Nachweis erbringen, dass das neue Produkt besser ist als Placebo, indem sie verkünden: „We have a new drug. It's non inferior to placebo." Der geballte Unsinn wird aber erst dann so richtig deutlich, wenn man sich vor Augen hält, dass die Autoren diesen Unsinn schreiben, die Gutachter diesen Unsinn tolerieren, der Herausgeber diesen Unsinn publiziert und die Leser des NEJM diesen Unsinn bei ihren Patienten auch noch anwenden. America first, oder?

Gut, dass laut „WIdOmonitor" nur 15 Prozent der deutschen Ärzte englischsprachige Journals tatsächlich nutzen. Aber wie kommen solche Publikationen überhaupt in solche Journals?

Das sage ich hier nicht, das kann man aber nachlesen in „Nature". Dort hat Professor David B. Allison* sehr gut beschrieben, wie in großen Journals gemogelt wird.

Was kostet eine Publikation in solch hochgerankten Journals?

Nichts. Aber wenn nach der Einreichung versucht wird, das Paper zurückzuziehen, kostet das bei manchen Journals bis zu 10.000 US-Dollar. Das steht alles in dem Allison-Paper wortwörtlich so drin. Bei diesen Preisen überlegt man sich schon, ob man von seinem knappen Forschungsgeld 10.000 Dollar ausgibt, bloß um den eigenen Schrott wieder rauszubekommen. Das ist schon Irrsinn.

Und bei Open Access?

Das kostet auch nichts, nur kommt da jemand ohne Geld nicht rein. Das bezahlen die hinter Studien stehenden Firmen über Anzeigen, meist aber über Sonderdrucke.

Das lohnt sich?

Und wie. Ich bin befreundet mit einem Senior Editor eines weltweiten Journals, den ich einmal gefragt habe, warum die solchen Blödsinn veröffentlichen. Seine Antwort: „Wenn ein Unternehmen für 2,5 Millionen Dollar Sonderdrucke bestellt, überlegt man sich das schon ..."

Herr Prof. Porzsolt, danke für das Gespräch. <<

Das Interview führte MVF-Chefredakteur Peter Stegmaier.

*Kommentar von David B. Allison, Andrew W. Brown, Brandon J. George & Kathryn A. Kaiser: „Reproducibility: A tragedy of errors", in „Nature" (3 February 2016)

Zitationshinweis

Stegmaier, P.: „Externe Evidenz ist genauso wichtig wie interne", in „Monitor Versorgungsforschung (03/17), S. 10-15; doi: 10.24945/MVF.03.17.1866-0533.2013

Prof. Dr. med. Franz Porszolt

Der 1946 geborene Porzsolt studierte von 1967 bis 1973 Humanmedizin an der Philipps-Universität in Marburg, war 1973-74 als Medizinalassistent am Universitätsklinikum Marburg tätig und arbeitete 1975-76 als Stipendiat der Deutschen Forschungsgemeinschaft am Ontario Cancer Institute in Toronto/Ontario. Von 1976 bis 1982 wurde er zum Facharzt für Innere Medizin, von 1982 bis 1985 zum Facharzt für Hämatologie und Internistische Onkologie ausgebildet. Danach arbeitete er als Leitender Oberarzt der Abteilung Innere Medizin III, anschließend als Geschäftsführender Oberarzt der Medizinischen Klinik des Universitätsklinikums Ulm. Er war zehn Jahre wissenschaftlicher Sekretär des Tumorzentrums Ulm und gründete 1992 die Arbeitsgruppe Klinische Ökonomik. Er ist Fellow der New York Academy of Medicine, hat rund 400 wissenschaftliche Originalarbeiten und Buchbeiträge verfasst, ist Gutachter der IAEA in Wien und wurde mit dem Deutschen IQ-Preis für Wissenschaft und Innovation ausgezeichnet. Aktuell ist er Leiter Versorgungsforschung an der Klinik für Allgemein- und Viszeralchirurgie des Universitätsklinikums Ulm sowie Vorsitzender des Vorstandes des von ihm initiierten Institute of Clinical Economics (ICE) e.V. in Ulm.