Wissenschaftliches Symposium „RCT – Bias-Kontrolle?“ des ICE e.V.

Über die Efficacy zur Effectiveness zum Value

Auf dem Fachkongress „Extrapolation 2018“, veranstaltet am 11. April 2018 von „Monitor Versorgungsforschung“ (MVF), wurde klar, dass Randomized Controlled Trials (RCT) zwar von allen Referenten als Goldstandard zum Nachweis der Wirksamkeit unter Idealbedingungen geschätzt werden. Unklar ist jedoch noch, mit welchen konsentierten Prozessen die in RCT gewonnenen Ergebnisse in die Realversorgung zu übertragen – zu extrapolieren – sind. Grund genug, um auf diesen ersten deutschsprachigen Fachkongress ein wissenschaftliches Symposium mit dem Titel „RCT – Bias-Kontrolle?“ folgen zu lassen. Es diskutierten PD Dr. Stefan Lange (Stellv. Leiter des IQWiG) mit dem Gesundheitsökonomen Prof. Dr. Jürgen Wasem und dem Begründer der Klinischen Ökonomik und des Institute of Clinical Economics (ICE) e.V., Prof. Dr. med. Franz Porzsolt.

>> „Goldenes Lamm, heilige Kuh oder einfach nur vernünftig?“ So überschrieb PD Dr. Stefan Lange, stellvertretender Leiter des Instituts für Qualität und Wirtschaftlichkeit im Gesundheitswesen (IQWIG) seinen Eröffnungsvortrag des Wissenschaftlichen Forums „RCT - Biaskontrolle?“ in Berlin, veranstaltet als Auftaktveranstaltung einer geplanten Reihe derartiger methodischer Foren vom Institute of Clinical Economics e.V. (ICE). Lange, der sich selbst ein „Faible für Rationalität“ zuschrieb, weshalb er beim IQWiG arbeiten würde, zitierte eingangs seines Vortrag den §2 Abs. 3 AM-NutzenV: „Der Nutzen eines Arzneimittels im Sinne dieser Verordnung ist der patientenrelevante therapeutische Effekt insbesondere hinsichtlich …“. Er hob hier den Begriff des „Effekts“ hervor, um die sich die gesamte frühe Nutzenberwertung nun einmal drehen würde.

Dieser Effekt, so Lange, der dafür „Wikipedia“ bemühte, sei nichts anderes als „eine durch eine bestimmte Ursache hervorgerufene Wirkung“, begründet auf einer Kausalität. Und diese sei nun einmal mit einem randomisierten Experiment darstellbar, wobei Randomisierung seinen Worten zufolge „nicht mehr, sondern (mitunter erheblich) weniger Aufwand“ bedeute, um zu aussagekräftigen Ergebnissen zu kommen, wenn es um Nutzen (oder Schaden) gehe.

Denjenigen, die als ein häufiges Argument kontra RCT behaupten würden, die Verblindung („gerne von Medizinprodukte-Herstellern bemüht“) sei unmöglich, oder auch die Seltenheit von Erkrankungen oder Ereignissen heranziehen, entgegnete er: „Der Verzicht auf eine randomisierte Zuteilung macht eine seltene Erkrankung nicht häufiger oder ein seltenes Ereignis nicht relevanter.“ Doch auch genau darum sei es „nicht vernünftig, wenn bereits eine Komponente der Ergebnissicherheit kompromittiert ist (Präzision), auch noch eine zweite (Qualität) einzuschränken“. Sein Zwischenfazit lautete daher: „Der Verzicht auf eine randomisierte Zuteilung ist bei seltenen Erkrankungen, kleinen Zielpopulationen oder seltenen Ereignissen genauso wenig rational wie bei häufigen Erkrankungen oder häufigen Ereignissen.“

Auch gegen den hilfsweisen Einsatz von Real World-Daten sprach sich Lange aus, nicht, weil sie generell schlechter seien, sondern weil es im Prinzip einzig und alleine auf die Begründung und Darstellung eines Effekts ankomme. Wenn man nun für eine bestimmte Fragestellung auf der einen Seite Ergebnisse eines RCT, auf der anderen Seite Daten aus dem nicht-randomisierten Alltag vergleichen wollte, handelt es sich laut Lange dabei um eine „Variation von mindestens zwei möglichen Effektmodifikatoren“, was ihn zur Kernfrage führt: „Wie sollen nun Unterschiede in Effektschätzungen interpretiert werden?“



v.li.: PD Dr. Stefan Lange (stellv. Leiter des IQWiG), Prof. Dr. med. Franz Porzsolt und Gesundheitsökonom Prof. Dr. Jürgen Wasem: RCT sind eben doch eine recht harte Nuss.

Daher sein Rat, weil ansonsten keine eindeutige Interpretation möglich sei: „Wer einen eventuellen Effekt-Modifikator („Alltag“ vs. „Idealbedingungen“) identifizieren möchte, tut nicht gut daran, gleichzeitig ein weiteres (Studien-) Merkmal zu modifizieren.“ Das liege aber auch daran, dass Daten aus Beobachtungsstudien – die sogenannte „Real World-Evidence“ – seiner Ansicht nach „immanent einem Selektionsbias unterliegen würden, „der durch kein statistisches Verfahren der Welt kontrolliert werden“ könne. Und das führt ihn wiederum zu seinem absolut eindeutigen, gleichsam kategorischen Abschlussfazit, das da lautet: „Es existiert (nahezu) kein rationales Argument, um für eine Nutzenbewertung auf nicht randomisierte Studien auszuweichen.“

Wenn man Langes Vortrag Revue passieren lässt, kann man sich des Eindrucks nicht erwehren, dass RCT zwar vernünftig, aber sicher auch eine Art „heilige Kuh“ sind, weil der stellvertretender Leiter des IQWiG auf keinen einzigen Punkt einging, der den „Goldstandard RCT“ in Frage stellen könnte.

Das geschah erst nach einem Vortrag von Prof. Dr. rer. pol. Jürgen Wasem, dem Lehrstuhlinhaber für Medizin-Management an der Universität Duisburg-Essen, der als Ökonom nach eigenem Bekunden mit allen verfügbaren – wenn zur Fragestellung passenden – Instrumenten arbeiten würde, frei nach dem Motto: „Der Gesundheitsökonom ist genügsam und nimmt, was er kriegen kann, um es für seine Zwecke weiter zu verarbeiten.“ Für den hoch angesehenen und ihn vielen wichtigen Funktionen im Gesundheitssystem aktiven Ökonomen gelte nach wie vor ein Statement, das einer vom BMG, im Jahre 2007 noch unter Gesundheitsministerin Ulla Schmidt eingesetzten Expertengruppe (Jöckel, Kohlmann, Raspe, Wasem) stammt. Und das da lautet: „Nutzen setzt den Nachweis der Wirksamkeit voraus. Randomisierte klinische Studien (RCT) stellen den Goldstandard für den Nachweis von Wirksamkeit und Nutzen dar. Allerdings kann es notwendig werden, für weitere Nutzenbelege auch andere Studientypen heranzuziehen. Bei der Auswertung und Interpretation von Studien niedrigerer Evidenzgrade ist eine besonders sorgfältige Prüfung auf systematische Verzerrungen notwendig.“

Seiner Meinung nach aber wäre es, sofern es bei Erstattung und Preisbildung um die Zahlungsbereitschaft für Zusatznutzen geht, vorzuziehen, dass dem „realen Geld“ auch „realer Zusatznutzen“ gegenüber stünde, das heißt, „sich nicht nur Efficacy im künstlichen RCT-Design, sondern auch Effektivität in der Versorgungswirklichkeit gezeigt hat“. Aufgrund der Komplexität und He-

terogenität des Versorgungs- und Krankheitsgeschehens setzt dies seiner Meinung nach „auch den Einsatz von Non-RCT voraus".

Generell nennt er jedoch auch das Problem, dass sowohl nicht-randomisierte (mindestens) zweiarmige Studien, in denen die Probanden im Kontrollarm prospektiv eingeschlossen werden, als auch einarmige Studien mit retrospektivem Hinzumatchen von Kontrollarmen (z.B. aus Routinedaten generiert) an der Problematik des Selbstselektionsbias leiden würden, so dass seiner Ansicht nach das mögliche Ausmaß des Bias nur qualitativ abgeschätzt und kritisch diskutiert werden müsste.

Das sieht auch Prof. Dr. med. Franz Porzsolt, Vorsitzender des von ihm gegründeten Institute of Clinical Economics e.V. (ICE), so, der mit seinem Vortrag eine „offene Diskussion über die Risiken von Bias" anstoßen will. Seiner Meinung ist diese Diskussion dringend erforderlich, weil nur so drei Kernfragen von Archie Cochrane und Austin Bredford-Hill beantwortet werden könnten: „Can it work?" (Efficacy), „Does it work?" (Effectiveness) und „Is it worth it? (Value).

Um sich diesen Fragen zu nähern, müsse man offen über Stärken, aber vor allem auch Grenzen von RCT sprechen dürfen. Zu den Schwächen zähle man gemeinhin, dass RCT keinen Anspruch auf Beschreibung der „Effectiveness" unter Alltagsbedingungen erheben könnten, aber dass dafür durch ihr Studiendesign eine „ähnliche Verteilung bekannter und unbekannter Risikofaktoren der untersuchten Gruppen" und eine „valide Beschreibung der „Efficacy" unter Idealbedingungen (Ausschluss von Störfaktoren) möglich sei.

Dahinter macht Porzsolt gleich mal ein ganz dickes Fragezeichen und stellt damit einige, bisher als schiere Wahrheit hingenommene Lehrmeinung auf den Kopf. Seiner Meinung nach ist nicht nur die externe Validität fraglich, weil hier Ausschlusskriterien, die Arzt-Präferenz an Studien überhaupt teilzunehmen und auch die Stärke der Präferenzen der Probanden hineinspielen würden, sondern auch die selten hinterfragte interne Validität. Hierzu zählt für Porzsolt, der Internist und Onkologe als auch den klinischer Ökonom ist, sowohl die Nichtbeachtung „bekannter" Risikofaktoren, als auch die Effekte unbekannter Risikofaktoren, die durch die Anwendung der Randomisation bekanntlich aus der Welt geschafft werden sollen.

Doch: Ist das richtig? Seine Antwort folgte stante pede: „Nein, das ist falsch!" Das liege vor allem darin begründet, dass die sogenannte „Table One" (in der bei allen RCT alle Ein- und Auschlusskriterien gelistet werden müssen) nur bekannte Risikofaktoren aufgenommen werden, weil man durch das randomisierende Studiendesign sich in einer trügerischen Sicherheit wähnt. Begründet ist dies in der Annahme, dass durch die Zufallsverteilung als solcher unbekannte Risikofaktoren auf beide Untersuchungsarme gleich verteilt seien. Doch gelte, so Porzsolt: „Je kleiner die Studie, um so größer das Risiko der unbekannten Risikofaktoren."

Bisher ist das absolute Terra incognita, weil niemand die, sich einzig und alleine aus der Statistik ergebenden Grenzwerte kennt. Darum hat Porzsolt von der Mathematikerin Christel Weiß aus Mannheim rechnen lassen, wie groß eine Studie überhaupt sein muss, dass unter der Annahme von nur 10 unbekannten Risikofaktoren alle unbekannten Faktoren auf die Studienarme gleich verteilt sind. Die rein mathematisch-statistische Antwort, aufbauend auf absolut konservativen Annahmen der Risikofaktor-Kategorien: dichotome statt kontinuierlich verteilte sowie von einander unabhängige statt abhängige Variablen, zudem ein zulässiger Grenzwert von 5% (statt 1%) Abweichung für eine anzunehmende Gleichverteilung.

Das Ergebnis: Eine tolerierte 5%ige Irrtumswahrscheinlichkeit wird erst bei 0,00328 unterschritten, was erst bei „n über 1.000" der Fall ist. Will auf gut deutsch heißen: Alle RCT unter einer Fallzahl von 1.000 Probanten können alleine schon aus mathematisch-statistischen Gründen unbekannte Risikofaktoren gar nicht gleich verteilen. Da aber die allermeisten veröffentlichten RCT eine geringere Fallzahl als 1.000 aufweisen, ist nach Porzsolt „eben nicht zu gewährleisten, dass unbekannte Risikofaktoren gleich verteilt sind".

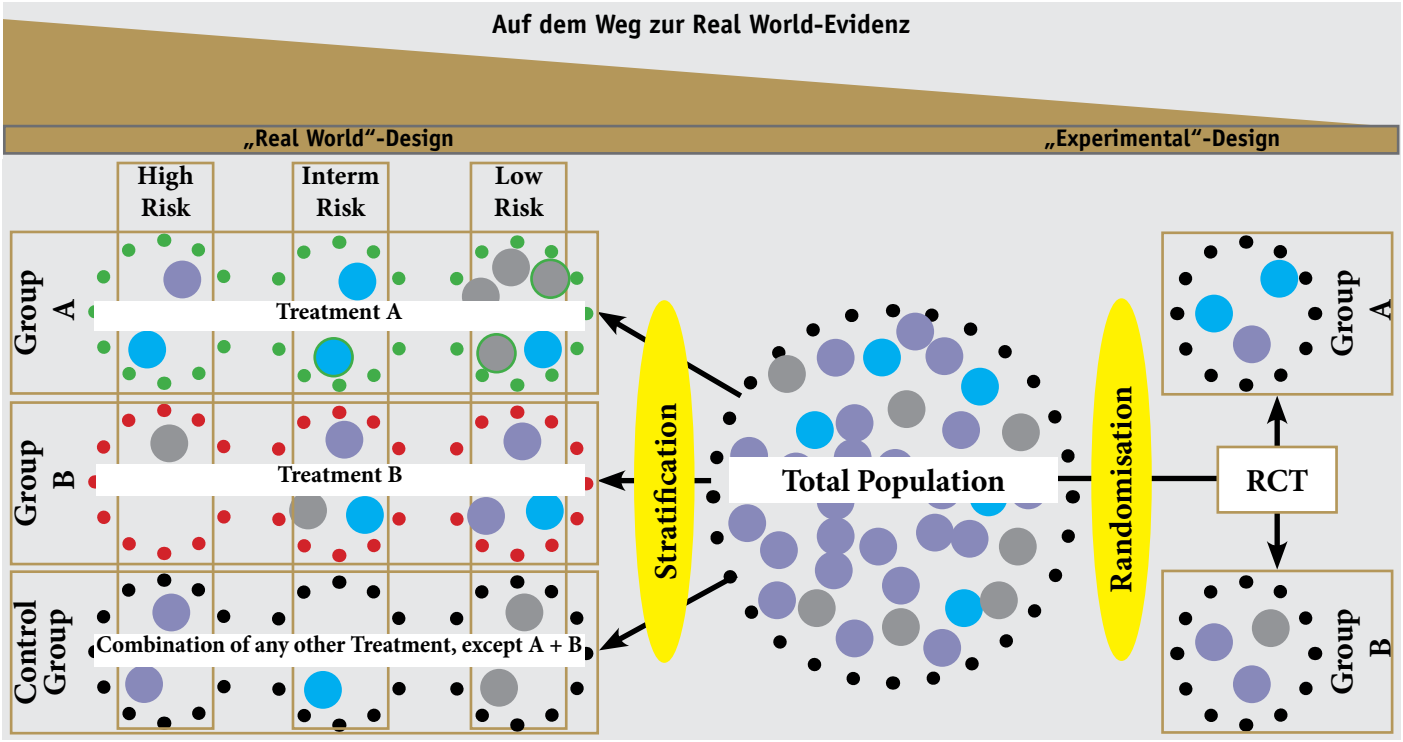


Abb.1.: Statt randomisierter Mixed-Risk-Groups sollen Gruppen mit gleichen Endpunkt-spezifischen Ausgangsrisiken verglichen werden, da ein PCT die Ergebnisse getrennt für jede endpunkt-adaptierte Risikoklasse beschreiben kann. Aus Vortrag: Prof. Dr. Franz Porzsolt, Institute of Clinical Economics (ICE): „Diskussion über die Risiken von Bias in Randomized (RCT) und Pragmatic Controlled Trials (PCT) am 25. April 2018 in Berlin. Quelle: Porzsolt F, Rocha NG, Toledo-Arruda AC, Thomaz TG, Moraes C, Bessa-Guerra TR, Leão M, Migowski A, Araujo de Silva AR, Weiss C.: „Efficacy and Effectiveness Trials Have Different Goals, Use Different Tools, and Generate Different Messages", in: „Pragmatic and Observational Research" 2015; 6:47-54. DOI http://dx.doi.org/10.2147/POR.S89946

Zitationshinweis

Stegmaier, P.: „Über die Efficacy zur Effectiveness zum Value“, in: „Monitor Versorgungsforschung“ (03/18), S. 28-30, doi: 10.24945/MVF.03.18.1866-0533.2078

Über den Unterschied zwischen Efficacy unter Idealbedingungen und Effectiveness im Alltag

„Versorgung und Forschung müssen nicht unbedingt miteinander etwas nichts zu tun haben. Wer versorgt, möchte ein Gesundheitsproblem seines Patienten lösen; präferiert deshalb eine der Versorgungsoptionen und hat kein großes Interesse, neues Wissen zu generieren. Wer hingegen forscht, möchte neues Wissen generieren und sollte darum keine der untersuchten Optionen präferieren. Der Aspekt der Patientenversorgung muss darum nachrangig sein, um den Prozess des Erkenntnisgewinns nicht zu beeinflussen. Klinische Wissenschaftler haben das Problem, beiden Zielen gerecht zu werden.

In meiner eigenen Entwicklung vom Kliniker zum Wissenschaftler hat mir einer meiner Patienten die Frage gestellt, ob ich tatsächlich von der Wirksamkeit und dem Nutzen der Kontrolltherapie überzeugt bin. An diesem Tage kamen meine ersten Zweifel auf, ob ich aus ethischen Gründen noch berechtigt bin, meine klinische Karriere als leitender Oberarzt unseres Klinikums in Ulm fortzusetzen.“

Prof. Dr. Franz Porzsolt

	Efficacy wird gemessen unter Idealbedingungen	**Effectiveness wird gemessen unter Alltagsbedingungen**
Eine Intervention wird mit dem Ziel angewandt …	… ein Humanexperiment durchzuführen, um neue Erkenntnisse zu gewinnen. Der Versorgungsaspekt ist unbedeutend.	… die Patienten gemäß des Versorgungsauftrags zu versorgen; der Erkenntnisgewinn ist unbedeutend.
Respekt vor Autonomie des Patienten …	… wird eingeschränkt, weil für den Erkenntnisgewinn auf eine essentielle ärztliche Leistung, die Auswahl der „bestmöglichen Intervention“, zu verzichten ist.	… wird gewährleistet, weil nicht auf die essentielle Leistung des Arztes verzichtet werden muss, die „bestmögliche Intervention“ für den Patienten auszuwählen.
Die Legitimation zur Anwendung der Intervention …	… ist durch das Votum einer Ethikkommission gegeben	… ist durch den Abschluss eines Versorgungsvertrages oder durch einen Notfall gegeben (§§ 95, 95a und 95c SGB V - §§ 3 ff., 16b, 18 ff. Ärzte-ZV – § 17 BMV-Ä – Bedarfsplanungs-Richtl. des G-BA).
Anwendung einer klinisch ungeprüften Intervention	ist die Regel.	Ist die Ausnahme, weil der Nachweis der Efficacy bereits vorliegt.
Der Nutzen der Intervention für den Patienten …	… wird geprüft.	… ist bereits nachgewiesen.

Tab.1: Idealbedingungen oder Alltagsbedingungen? Ideal Study Conditions oder Real World Conditions? Efficacy oder Effectiveness?

Um dennoch einen Lösungsansatz diskutieren zu können, plädiert Wissenschaftler Porzsolt für ein aufeinander aufbauendes Stufenmodell. In der für ihn nach wie vor „allerersten und wichtigsten Stufe“ RCT muss erst einmal nachgewiesen werden, dass für eine Innovation unter Idealbedingungen die Wirkung (Efficacy) nachgewiesen werden kann. Geschieht das nicht, ist jede weitere Beschäftigung mit dem Thema unsinnig.

Kann aber eine Innovation in Hürde 1 (RCT) eine Wirkung („can it work“ – „Efficacy“) unter Idealbedingungen nachweisen, muss die nächste Stufe 2 folgen, welche der Frage „does ist work – Effectiveness“ unter Alltagsbedingungen beantwortet. Dieser zweite Nachweis ist Porzsolts Überzeugung nach enorm bedeutend, weil unter Alltagsbedingungen nicht alles funktioniert, was unter Laborbedingungen klappt und umgekehrt (siehe dazu Tab. 1): „Manches im Alltag läuft sogar besser als im Labor.“ Genau darum müsste in Stufe 2 die Frage des „does it work“ erforscht werden: „Hält die Innovation auch im Alltag das, was sie im Labor verspricht?“

Efficacy einer- und Effectiveness andererseits müsse man nach Porzsolt nicht nur als zwei Seiten einer Medaille, sondern als zwei unterschiedliche Dimensionen verstehen. Der Grund: Man könne mit dem Ergebnis der einen Dimension (z.B. Efficacy) nicht auf das der anderen (z.B. Effectiveness) schließen – was umgekehrt ebenso gilt. Porzsolts Behauptung: „Wenn man Efficacy misst, ist das Effectiveness-Ergebnis unvorhersehbar: Es kann besser, aber auch schlechter sein, was wiederum mit den unvorhersagbaren Risikoverteilungen zu tun hat.“

Ist nun das Effectiveness-Ergebnis positiv, muss nach Porzsolts Überzeugung auch noch eine Stufe 3 folgen, beantwortend die Frage: Wenn nun eine Intervention unter Idealbedingungen wirkt (Stufe 1, d.h. wenn alle Confounder eliminiert sind) und einen ähnlichen Effekt auch unter Alltagsbedingungen zeigt (Stufe 2, d.h. wenn keiner der Confounder eliminiert ist), dann stellt sich die Frage, ob dieser (statistisch signifikante Effekt) vom Patienten und der Gesellschaft auch als Mehrwert empfunden wird (is it worth it – „value“)!

Die zweite Frage („does it work?“) richtet sich an die Alltagsbedingung und kann deshalb nach Porzsolts fester Meinung „nicht im Labor“ beantwortet werden. Deshalb nennt er das dafür einsatzbereite Studiendesign ein „Pragmatic Trial“; und weil dieses Trial eine Kontrollgruppe enthält, wird es „Pragmatic Controlled Trial“ (kurz: PCT) genannt. Dieses PCT sei ein deskriptives Verfahren, um Effekte, die unter Idealbedingungen in einem RCT statistisch als Efficacy nachgewiesen wurden, auch unter Alltagsbedingungen – aber ebenfalls mit statistischen Methoden – als Effectiveness zu bestätigen.

Im dritten Schritt muss dann aber auch die Frage beantwortet werden, ob der im Alltag tatsächlich auftretende Effekt (z.B. signifikante Verlängerung des Schlafs durch ein neues Schlafmittel) auch als „klinisch relevant“ oder „wertvoll“ („is it worth it“) empfunden wird. Diesen Nachweis kann man – so Porzsolt – weder mit einem RCT, noch mit einem PCT erbringen: „Ob etwas den Aufwand wert ist, wird jeder individuelle Patient für sich selbst abschätzen, ohne einen Rechner zu benutzen. Aber auch jeder erfahrene Manager wird die Frage für sein Unternehmen ohne Papier und Bleistift beantworten können.“ Dennoch gebe es Wissenschaftler, die diese „Abschätzung“ nicht akzeptieren und einen Rechenalgorithmus fordern, dessen Richtigkeit kaum jemand prüfen kann.

Insgesamt ergeben die drei von ihm vorgestellten Evaluierungs-Schritte gemeinsam ein klares Konzept:

1. dem RCT,
2. dem PCT und
3. einem ökonomischen Bewertungsverfahren oder einer Schätzung.

Mit diesen drei unterschiedlichen, aufeinander aufbauenden (nicht jedoch sich gegenseitig kannibalisierenden) Verfahren könnten die drei 40 Jahre alten Kernfragen der Gesundheitsforschung beantwortet werden. Porzsolt: „Diese konsekutiven Verfahren beschreiben zunächst das Prinzip, nach welchen Schritten eine Nutzenbewertung im Rohentwurf aussehen könnte. Damit sind wir aber bereits einen erheblichen Schritt weiter als ohne sie.“ Ob das wirklich so ist, muss erst noch ebenso offen wie streitfreudig diskutiert und vor allem konsentiert werden. Hier bittet Porzsolt, doch härter an der Wissenschaft zu bleiben, statt allzu schnell auf eine politisch-bewertende Diskus-sionsebene abzugleiten. Daher möge jeder selbst entscheiden, welche der vorgetragenen Aussagen auf wissenschaftlichen Grundlagen und welche auf politischen Bewertungen beruhen, die erneut beim 2. ICE-Diskussionsforum im Herbst 2018 aufeinandertreffen können. <<

von: MVF-Chefredakteur Peter Stegmaier