ROBERT KOCH INSTITUT

Beiträge zur
Gesundheitsberichterstattung
des Bundes

Evaluation komplexer Interventionsprogramme in der Prävention: Lernende Systeme, lehrreiche Systeme?

Gesundheitsberichterstattung des Bundes

Beiträge zur
Gesundheitsberichterstattung
des Bundes

Evaluation komplexer Interventionsprogramme in der Prävention: Lernende Systeme, lehrreiche Systeme?

Robert Koch-Institut, Berlin 2012

Bibliografische Information Der Deutschen Bibliothek
Die Deutsche Bibliothek verzeichnet diese Publikation
in der Deutschen Nationalbibliografie.

Herausgeber
Robert Koch-Institut
Nordufer 20
13353 Berlin

Bayerisches Landesamt für Gesundheit und Lebensmittelsicherheit
Veterinärstraße 2
85764 Oberschleißheim

Konzept und Redaktion
Dr. Joseph Kuhn
Bayerisches Landesamt für Gesundheit und Lebensmittelsicherheit
Dr. Thomas Ziese, Dr. Thomas Lampert
Robert Koch-Institut

Zitierweise
Robert Koch-Institut, Bayerisches Landesamt für Gesundheit und Lebensmittelsicherheit (Hrsg) (2012) Evaluation komplexer Interventionsprogramme in der Prävention: Lernende Systeme, lehrreiche Systeme? Beiträge zur Gesundheitsberichterstattung des Bundes. RKI, Berlin

Abonnentenservice
E-Mail: gbe@rki.de
www.rki.de/gbe
Tel.: 030-18754-3400
Fax: 030-18754-3513

Grafik/Satz
Gisela Winter
Robert Koch-Institut

Druck
Ruksaldruck, Berlin

ISBN
978-3-89606-215-4

Inhaltsverzeichnis

Vorwort

Das Robert Koch-Institut hat in vielen Public Health-Bereichen zentrale Aufgaben beim Erkennen und Bewerten von Entwicklungen im Bereich Gesundheit. Neben den gut etablierten Surveillance-Systemen bei den Infektionskrankheiten verfügt das RKI mit dem Gesundheitsmonitoring und der Gesundheitsberichterstattung über ein, auch im internationalen Vergleich, gut ausgebautes Surveillance-System zu Gesundheitsrisiken, Gesundheitsstatus und Gesundheitsoutcome. Mit diesen Instrumenten können Veränderungen im Gesundheitsverhalten und im Krankheitsspektrum im Verlaufe von Interventionen erkannt und analysiert werden. Da das Gesundheitsmonitoring im Hinblick auf Public Health-Aufgaben konzipiert wurde, spielt der Bereich Prävention eine tragende Rolle. Zusätzlich wurden für nationale Aktionsprogramme wie beispielsweise gesundheitsziele.de oder IN FORM zusätzliche Instrumente bzw. vertiefende Elemente sowohl in das Gesundheitsmonitoring als auch in die Gesundheitsberichterstattung des Bundes integriert. Das RKI kann damit für Interventionsprogramme im Bereich der Prävention eine Globalbewertung bereitstellen, die ein zentrales Element einer Evaluation darstellt.

Die Evaluation von komplexen Interventionen geht jedoch deutlich darüber hinaus, und erfordert die Beteiligung aller Akteure der jeweiligen Prozesse. Je breiter das Zielespektrum angelegt ist, desto mehr Akteure spielen eine aktive Rolle bei der Durchführung der Programme und umso differenzierter muss die Betrachtung der Prozesse, Strukturen und Outcomes bei der Ergebnisbewertung erfolgen. Je komplexer Interventionen angelegt sind, desto komplexer ist meist auch ihre Evaluation, die mehr ist als die summierte Evaluation von Einzelmaßnahmen. Vielfach fehlen jedoch zurzeit noch klare Rahmensetzungen über Art und Umfang von Evaluationen entsprechender Aktionsprogramme.

Komplexe Public Health-Maßnahmen sind von ihrer Natur her interdisziplinär angelegt und beinhalten unterschiedliche Public Health-Bereiche wie Forschung, Anwendungsebene, Koordination und Finanzierung. Entsprechend müssen auch die Evaluationen interdisziplinär angelegt sein und gemeinsam mit den beteiligten Akteuren erarbeitet werden. Um den Diskurs über die gemeinsame Entwicklung von Evaluationskonzepten weiter zu führen, wurde 2011 in enger Zusammenarbeit mit dem Bayerischen Landesamt für Gesundheit (unser besondere Dank geht dabei an Dr. Joseph Kuhn) der hier dokumentierte Workshop zur Evaluation komplexer Interventionen durchgeführt. Die vorgestellten Konzepte und methodischen Arbeiten sollen das breite Spektrum unterschiedlicher Ansätze darstellen, die Gemeinsamkeiten aufzeigen und so zur Entwicklung von gemeinsamen Standards zur Evaluierung solcher Interventionsprogramme beitragen.

Wir wünschen Ihnen eine spannende Lektüre, und freuen uns auf konstruktive Diskussionen bei der Weiterentwicklung von Evaluationskonzepten von Interventionsprogrammen auf nationaler und regionaler Ebene.

Dr. Bärbel-Maria Kurth
Leiterin der Abteilung für Epidemiologie
und Gesundheitsberichterstattung

Dr. Thomas Ziese
Leiter des Fachgebietes
Gesundheitsberichterstattung

Einführung ins Thema: Komplexe Interventionen – komplexe Evaluationen?

Joseph Kuhn, Thomas Lampert, Thomas Ziese

Zu wissen, ob eine Intervention wirksam ist oder nicht und ob der Nutzen einer Intervention ihre Risiken überwiegt, ist im Gesundheitswesen in vielen Bereichen von höchster Bedeutung. In der kurativen Medizin ist diese Bedeutung durch die Herausbildung der »Evidenzbasierten Medizin« seit einiger Zeit in besonderer Weise unterstrichen worden. Anspruch der Evidenzbasierten Medizin (EBM) ist es, das best verfügbare Wissen für klinische Entscheidungen zusammenzuführen: »EBM ist der gewissenhafte, ausdrückliche und vernünftige Gebrauch der gegenwärtig besten externen, wissenschaftlichen Evidenz für Entscheidungen in der medizinischen Versorgung individueller Patienten« (Sackett et al. 2012, S. 14). Dabei wird die klinische Erfahrung nicht ausgeblendet, sondern sie soll im Lichte der wissenschaftlichen Befunde zur Geltung kommen. Dieser Ansatz wurde schon nach kurzer Zeit auch auf Public Health-Maßnahmen übertragen. Brownson et al. (1999, S. 87) definieren Evidence-Based Public Health beispielsweise als "the development, implementation, and evaluation of effective programs and policies in public health through application of principles of scientific reasoning including systematic uses of data and information systems and appropriate use of program plannung models". Auch hier geht es darum, der praktischen Erfahrung wissenschaftliche Evidenz zur Seite zu stellen, in diesem Fall auf der Ebene von bevölkerungsbezogenen Interventionen.

Evidenzbasierung und Evaluation stehen dabei gewissermaßen in einem komplementären Verhältnis zueinander. Evidenzbasierung verweist auf vorhandenes Wissen, Evaluation verweist auf nicht vorhandenes Wissen. Sie soll Evidenz erst schaffen, ex post, anders als z. B. ein health impact assessment (Kemm 2003), das ex ante die gesundheitlichen Folgen einer Intervention abschätzen soll.

Die Unterscheidung zwischen wirksamen und nicht wirksamen Maßnahmen zieht – das ist der Sinn der Sache – eine Selektion von Maßnahmen nach sich. Die konkrete Umsetzung dieser Selektion kann unterschiedlich ausfallen, von der qualitätssichernden Anpassung an die als wirksam erkannte Form der Durchführung bis hin zum völligen Verzicht auf eine als unwirksam erkannte Maßnahme. Gerade im Public Health-Bereich kann dies empfindliche wirtschaftliche oder politische Konsequenzen für die Institutionen nach sich ziehen, die für die jeweiligen Interventionen verantwortlich zeichnen. Evidenzfragen sind nie nur wissenschaftliche Fragen, sie werden immer in einem gesundheitspolitischen Kontext gestellt und beantwortet.

In der Prävention gilt dies in besonderem Maße. Der Prävention werden große gesundheitliche und ökonomische Potenziale zugeschrieben. Sie soll gesundheitliche Risiken eindämmen, die mit unserer Lebensweise einhergehen, sie soll der Verbreitung chronischer Erkrankungen entgegenwirken, die kurativ nur bedingt zu beeinflussen sind, sie soll eine Antwort geben auf die Alterung der Gesellschaft und die in diesem Zusammenhang befürchteten Kostensteigerungen im Gesundheitswesen. Das sind große Erwartungen und die Frage, mit welchen Interventionen sie zu erfüllen sind, ist mehr als berechtigt. Trotz unverkennbarer Fortschritte in der Präventionsforschung weisen in der Praxis viele Kampagnen und Programme erhebliche Evidenzschwächen auf (vgl. Kliche et al. 2006). An Beispielen mangelt es nicht: Was genau eine »gesunde Ernährung« ausmacht und welchen Nutzen konkrete Ernährungsempfehlungen haben, ist nur zum Teil bekannt, ebenso, wie man Adipositas oder einen riskanten Alkoholkonsum bei Jugendlichen verhindert, oder welchen Nutzen manche Präventionskurse zur Vermeidung von Rückenbeschwerden im Betrieb haben. Unter Evaluationsgesichtspunkten werden hier aber sogar noch vergleichsweise einfache Fragen gestellt: Dies gilt nicht nur für die Fragen nach der Prozessqualität, etwa, wie viele Menschen der Zielgruppe erreicht wurden. Auch im Hinblick auf die Ergebnisqualität, die Wirksamkeit der Maßnahme, kann man die Fragestellung hier so einzäunen, dass die Wirksamkeit einer spezifischen Intervention in einer definierten Population mit Blick auf einen einzelnen Endpunkt zu untersuchen ist. Welche Wirksamkeit hat eine bestimmte Ernährungsberatung, kombiniert mit einem bestimmten

Bewegungsangebot bei jüngeren Frauen der unteren Mittelschicht auf den Body-Mass-Index oder das Hip-Waist-Ratio in einem halben Jahr oder in einem Jahr. Je spezifischer die Frage formuliert wird, desto weniger kann man zwar mit der Antwort in der Präventionspraxis anfangen, aber dieses Problem hat man in anderen Bereichen auch und die Beantwortung vieler solcher spezifischer Fragen hilft irgendwann dann doch bei praktischen Problemen.

Dieses inkrementelle Wachstum des Wissens scheint jedoch kaum weiterzuhelfen, wenn es um »komplexe Interventionen« geht. Auch dieser Begriff hat seit einiger Zeit Konjunktur und umfasst sehr heterogene Interventionstypen. In der Literatur werden unter »komplexen Interventionen« solche mit mehreren interagierenden Komponenten verstanden und verschiedene Komplexitätsdimensionen, z. B. hinsichtlich des Outcomes, der Zielgruppen oder der Durchführenden, angesprochen (vgl. z. B. Craig et al. 2008, S. 979 f.).

Oft wird in der Diskussion um komplexe Interventionen nicht zwischen der Komplexität der Interventionen selbst, der Komplexität des Systems, in das interveniert wird, und der Komplexität der Beziehungen zwischen Intervention und System unterschieden. Ein didaktisch klar strukturierter Rückenkurs, der in einem Betrieb etabliert werden soll, ist eine einfache Intervention, trifft aber auf ein höchst komplexes System – den Betrieb mit seiner sozialen Vielfalt und Dynamik. Ein verschriebenes Arzneimittel – eine im Zuge des Zulassungsverfahrens hochgradig standardisierte Intervention – trifft ebenfalls auf ein höchst komplexes System: Menschen mit all ihren individuellen und situativen Besonderheiten. Zudem werden Arzneimittel nicht einfach »eingenommen«, sie stehen im Kontext einer Behandlungssituation, zu der viele andere Komponenten gehören, die den Heilerfolg beeinflussen, etwa die ärztliche Beratung oder die Beratung in der Apotheke. Der klinische Nutzen eines Arzneimittels ist also nicht allein eine Eigenschaft des Arzneimittels, sondern hängt von der ganzen Komplexität der Versorgungsrealität ab. Zurecht wird zwischen efficacy und effectiveness unterschieden. Diese Unterscheidung spiegelt das Bewusstsein um die Problematik der unterkomplexen Aussage »das Arzneimittel wirkt« wider. Die Wirksamkeit von Interventionen in komplexen Systemen lässt sich nicht angemessen beurteilen, wenn diese Systemkomplexität nicht berücksichtigt wird. Dies kann im Zuge einer Evaluation auch dazu führen, dass aus der »einfachen Intervention« differenziertere Vorgehensweisen entwickelt werden, z. B. für verschiedene Subgruppen. Die individualisierte Medizin ist eine solche Rückwirkung angesichts genetischer und metabolischer Diversität. Daran wird zugleich die enge Verbindung der Berücksichtigung von Kontextfaktoren und der Transferierbarkeit von Interventionen von einem Kontext in einen anderen deutlich.

Eine »Gesundheitsberatung«, die je nach aktuellem Beratungsbedarf Themen der Ernährung, der Bewegung, des beruflichen Umfelds und anderes aufgreift, wäre dagegen wohl schon von der Intervention selbst her »komplex« zu nennen. Nicht nur, weil sie im Detail nicht standardisiert ist, sondern z. B. auch, weil sich ihre Komponenten gegenseitig beeinflussen können. Auch hier muss die Evaluation mit dieser Komplexität umgehen, von der Definition geeigneter Endpunkte bis hin zur Berücksichtigung der Wechselwirkungen der einzelnen Komponenten. Und als »komplex« sind sicher auch gemeindeorientierte Interventionen zu bezeichnen, die aus ganzen Programmbündeln bestehen: komplexe Interventionen in komplexen Systemen. Gängige confoundersensitive Evaluationsverfahren sind bei gemeindeorientierten Interventionen oft aufgrund der Nichtverfügbarkeit von Vergleichsgemeinden, von Fällen und Kontrollen, kaum einsetzbar und die Möglichkeit zur Randomisierung gemeindeorientierter Interventionen gehört in der Evaluation eher zu den seltenen Ausnahmen (vgl. dazu auch Leyland 2010).

»Komplex« werden häufig auch Interventionen, die vor der Implementation nicht hinreichend konkretisiert wurden, was Maßnahmen, Zielgruppen und andere Merkmale angeht, und sich erst im Verlauf der Durchführung strukturell ausgestalten: Komplexität durch Veränderung in lernenden Systemen. Diese Offenheit ist manchmal unvermeidlich, manchmal durch ein reflektiertes Interventionsdesign zumindest reduzierbar (Campell et al. 2000; Craig et al. 2008), manchmal durch eine Aufklärung von Kontextfaktoren sowie tatsächlich durchgeführten Interventionen und die Reformulierung von Wirkungsannahmen auch nachträglich evaluierbar (Pawson et al. 2005).

In der vorliegenden Dokumentation geht es nicht darum, dieses Feld insgesamt begrifflich zu

ordnen und auf seine evaluativen Besonderheiten zu untersuchen. Vielmehr soll hier ein besonderer Typ von komplexen Interventionen betrachtet werden: Präventionsorientierte Interventionen, die mehrere inhaltliche – u. U. wechselwirkende – Dimensionen beinhalten, die von verschiedenen Akteuren auf unterschiedlichen Programmebenen umgesetzt werden, aber dennoch unter einem gemeinsamen strategischen Dach gebündelt sind und sich damit als »ein Programm« oder »eine Strategie« darstellen und legitimieren müssen.

Konkret geht es um große politisch initiierte Präventionsprogramme mit einem dezidierten Public Health-Anspruch:

- die Präventionsmaßnahmen der gesetzlichen Krankenversicherung nach § 20 SGB V,
- die Nationalen Gesundheitsziele,
- die Gemeinsame Deutsche Arbeitsschutzstrategie,
- den Nationalen Aktionsplan »IN FORM«,
- die Gesundheitsinitiative Gesund.Leben.Bayern.,
- die Netzwerke Gesunde Kinder im Land Brandenburg,
- den Fonds Gesundes Österreich.

Ein Kernproblem, das dabei aufgeschlossen werden soll, lässt sich in der Frage auf den Punkt bringen: Brauchen komplexe Interventionen dieser Art auch komplexe Evaluationen? In vielen Bereichen der Medizin, vor allem in der Arzneimittelforschung, ist der randomisierte klinische Versuch, das RCT, der Goldstandard bei Fragen der Wirksamkeit. Ein RCT ermöglicht eine gute Confounderkontrolle. Manche Autoren, z. B. Lüngen et al. (2009, S. 106), fordern daher, es »sollte auch bei Präventionsprogrammen auf das hochwertige Design von RCTs zurückgegriffen werden.« Bei vielen Praktikern bedingt die Anwendung des Etiketts »komplex« auf eine Intervention jedoch nahezu reflexartig die Ablehnung eben dieses Evaluationsverfahrens. Die Diffusität des Komplexen und die Klarheit des RCT scheinen nicht vereinbar. Vielmehr scheint eine komplexe Intervention auch eine komplexe Evaluation nach sich zu ziehen. Die Frage danach, was eine »komplexe Evaluation« sei, kommt allerdings häufig nicht über eine mehr oder weniger gut begründete Ablehnung des RCT hinaus und droht so schnell zu einer Chiffre für eine weniger anspruchsvolle Evaluation zu werden, nach dem Motto, wo es kompliziert wird, darf man nicht so strenge Maßstäbe anlegen (kritisch dazu z. B. Bödeker 2006 und Bödeker 2012 in diesem Band). Dennoch lassen sich ernste Einwände gegen die pauschale Anwendbarkeit von RCTs in Public Health-Zusammenhängen nicht von der Hand weisen, aus ethischem, methodischen und grundsätzlichen Überlegungen (siehe z. B. Victora et al. 2004; Kuhn 2007). Im Arbeitsschutz ist beispielsweise in allen Betrieben gleiches Recht anzuwenden, so dass einem RCT zu unterschiedlichen Aufsichtsstrategien schon aus ethischen und rechtlichen Gründen enge Grenzen gesetzt sind. Bei den oben genannten Programmen ist zudem zu bedenken, dass sie singulären Charakter haben, als Gesamtheit also gar nicht wiederholbar sind, so dass RCTs ohnehin nur auf mutmaßlich wirksame Einzelkomponenten anwendbar wären. Pragmatische Antwortversuche plädieren dafür, angesichts des unabweisbaren Bedarfs an Wirkungsbelegen einerseits und der Anwendungsschwierigkeiten von klassischen Studiendesigns multimodale Evaluationsverfahren einzusetzen, z. B. eine Kombination von quantitativen und qualitativen Methoden, von »objektiven« und »partizipativen« Verfahren oder von Selbst- und Fremdevaluation. Die Triangulation von Ergebnissen aus solchen Verfahrenskombinationen kann dann Ausgangspunkt differenzierter Theorie- und Methodenentwicklung sein (vgl. Kliche et al. 2006, S. 148).

Über methodische Fragen der Evaluation im engeren Sinne hinaus gilt es bei den genannten Programmen auch zu berücksichtigen, dass sie in hohem Maße politischen Zielen unterliegen, was die Rahmenbedingungen ihrer Evaluation nicht unerheblich beeinflusst, z. B. im Hinblick auf die Ergebniserwartungen unterschiedlicher politischer Stakeholder, bestehende öffentliche Rechenschaftspflichten (z. B. gegenüber Rechnungshöfen) oder die Diskrepanz der politisch für die Evaluation zur Verfügung gestellten Zeit und der wissenschaftlich dafür gewünschten Zeit (siehe zu solchen Aspekten auch Milton et al. 2011; Knieps 2009). Die von der Deutschen Gesellschaft für Evaluation formulierten »Standards für Evaluation« (DeGEval 2008) können hier hilfreiche Orientierungen geben.

In den folgenden Beiträgen werden diese und andere Probleme aufgegriffen. Ausgehend von einigen konzeptionellen Überlegungen und einem Blick in die Evaluationsansätze in anderen Politik-

feldern werden in einem zweiten Schritt die oben genannten Präventionsprogramme in ihren evaluativen Grundzügen dargestellt. Dies soll dabei helfen, anhand konkreter Praxisbeispiele Antworten auf typische Fragestellungen im Zusammenhang mit der Evaluation solcher Programme zu finden, z. B.

- was typische Strukturmerkmale der Evaluation komplexer Interventionen sind, z. B. die Kombination von Dachevaluation und komponentenspezifischen Einzelevaluationen, die Kombination von Fremd- und Selbstevaluation, die Betonung von prozessevaluativen Elementen gegenüber wirkungsevaluativen Elementen, die öffentliche Berichtlegung von Evaluationsergebnissen, die Existenz von Evaluationsbeiräten usw.,
- welche Folgen für die Evaluation die oft gar nicht oder nur schwach formulierten Wirkungshypothesen komplexer Interventionen haben,
- ob sich die Evaluationsansätze mit dem Verlauf der Interventionen verändern,
- ob die Evaluation komplexer Interventionen mehr ist als die Addition der Evaluation relevanter Komponenten dieser Interventionen oder
- ob es in komplexen Interventionen eigenständige Evaluationsdimensionen gibt, z. B. solche, die sich aus der Wechselwirkung zwischen Programmbestandteilen ergeben, aus den Förderstrukturen oder aus dem dezidiert politischen Hintergrund dieser Interventionen,
- ob es machbar und ratsam wäre, in großen politischen Präventionsprogrammen nur evidenzbasierte Interventionen durchzuführen, so dass die Ansprüche an die Wirkungsevaluation reduziert werden könnten und Aspekte der Struktur- und Prozessevaluation (Akzeptanz, Zielgruppenerreichung, Public Outreach etc.) in den Vordergrund rücken dürften,
- ob für die Evaluation solcher Programme spezifische Berichtsformate sinnvoll sind und inwiefern dabei Checklisten-Verfahren wie z. B. die TREND Statement Checklist (http://www.cdc.gov/trendstatement/docs/TREND_Checklist.pdf) hilfreich sein können,
- und last but not least, welche Rolle bevölkerungsbezogene Surveys bzw. Monitoringdaten der Gesundheitsberichterstattung für die Evaluation solcher Programme spielen.

Das Robert Koch-Institut und das Bayerische Landesamt für Gesundheit und Lebensmittelsicherheit hatten dazu am 5. Dezember 2011 einen Workshop organisiert, der das Thema aufschließen und einen Einstieg in die Diskussion erleichtern sollte. Dabei hat sich einerseits gezeigt, dass die Problemlagen, denen sich die Evaluation großer Präventionsprogramme ausgesetzt sieht, sehr ähnlich sind, von den politischen Rahmenbedingungen bis hin zu den fachlich-methodischen Fragen. Anderseits sind auch einige interessante Unterschiede zutage getreten, die zu diskutieren sich lohnt. Eine Übersicht über die Evaluationsansätze zeigt die folgende Tabelle 1:

Die Beiträge dieses Workshops werden hier, ergänzt um einige zusätzliche Texte, die die Workshop-Inhalte abrunden sollen, dokumentiert. Sabrina Scholz vom Bayerischen Landesamt für Gesundheit und Lebensmittelsicherheit sowie Gisela Winter vom Robert Koch-Institut danken wir für die redaktionelle Unterstützung bei der Erstellung des Bandes. Den Workshop-Teilnehmern und -Teilnehmerinnen danken wir für ihre Diskussionsbereitschaft und hoffen, mit der Dokumentation die eine oder andere Anregung für die weitere Entwicklung von Evidence-Based Public Health-Konzepten geben zu können.

Tabelle 1
Übersicht zur Evaluationsstruktur ausgewählter komplexer Präventionsprogramme

	Präventions-maßnahmen der GKV	Nationale Gesundheits-ziele	Gemeinsame Deutsche Arbeitsschutz-strategie	Nationaler Aktionsplan In Form	Gesundheits-initiative Gesund.Leben. Bayern.	Fond Gesundes Österreich
Gibt es eine gesetzliche Grundlage der Evaluation	(●) indirekt		●			
Gibt es eine Dachevaluation	●	(●)	●	●	●	●
Gibt es Projektevaluationen	●	(●)	●		●	●
Kann sich die Dachevaluation an konkreten Globalzielen orientieren	(●) Präventions-ziele der GKV		●		(●)	(●)
Gibt es eine systematische Verschränkung von Dach- und Projektevaluation			(●)		●	●
Gibt es Zwischenevaluationen	●	(●)	●		●	●
Gibt es einen Evaluations-beirat oder Steuerungskreis für die Dachevaluation		●	●			●
Gibt es ein Mentoring/ Beratung für die Projektevaluationen			(●)	● für den Schwerpunkt Aktionsbündnisse	(●)	(●)
Gibt es strukturierte Vorgaben für die Projektevaluation	(●) Instrumente und Indikatoren		(●)	● für den Schwerpunkt Aktionsbündnisse	●	● werden im Rahmen der Dachevaluation entwickelt
Gibt es ein Budget über 50.000 Euro für die Dachevaluation			●		● erste Programmphase	● für drei Jahre
Gibt es ein Budget über 50.000 Euro für die Projektevaluationen			(●) über die Länder für die Arbeitspro-gramme	● für den Schwerpunkt Aktionsbündnisse	●	● jeweils ca. 10 % der Fördersumme
Gibt es eine Fremdevaluation	●		● Dachevaluation	● Aktionsbündnisse und größere Projekte	● erste Programmphase	● Dachevaluation, Teilprojekte mit Budget ab 72.000 €, kleinere Projekte freiwillig
Gibt es eine Selbstevaluation	● für Projekte	(●)	(●) Arbeitsprogramme	(●)	●	● Teilprojekte mit Budget unter 72.000 €
Gibt es eine Strukturevaluation	●	(●)		(●)	●	
Gibt es eine Prozessevaluation	●	(●)	●	(●)	●	● Dachevaluation und Teilprojekte
Gibt es eine Ergebnis-evaluation	●		(●)	(●)	(●) erste Programmpha-se, aktuell nur bei Projekten	● Dachevaluation und Teilprojekte
Gibt es einen öffentlichen Evaluationsbericht	●		● in Vorbereitung	(●)		● Publikationen geplant

Legende: ● gegeben (●) mit Einschränungen gegeben

Literatur

Bödeker W (2006) Evidenzbasierung in Gesundheitsförderung und Prävention – Der Wunsch nach Legitimation und das Problem der Nachweisstrenge. In: Bödeker W, Kreis J (Hrsg) Evidenzbasierung in Gesundheitsförderung und Prävention. Wirtschaftsverlag NW, Bremerhaven

Brownson RC, Gurney JG, Land GH (1999) Evidence-Based Decision Making in Public Health. J Public Health Management Practice 5 (5): 86–87

Bundesministerium für Familie, Senioren, Frauen und Jugend (2000) Zielgeführte Evaluation von Programmen – ein Leitfaden. Eigenverlag, Berlin

Campell M, Fitzpatrick R, Haines A et al. (2000) Framework for design and evaluation of complex interventions to improve health. BMJ 321: 694–696

Craig P, Dieppe P, Macintyre S et al. (2008) Developing and evaluating complex interventions: the new Medical Research Council guidance. BMJ 337: 979–983

DeGEval (2008) Standards für Evaluation. Gesellschaft für Evaluation e. V., Hamburg

Kemm J (2003) Perspektives on health impact assessment. Bulletin of the World Health Organization 81 (6): 387

Kliche T, Koch U, Lehmann H et al. (2006) Evidenzbasierte Prävention und Gesundheitsförderung. Bundesgesundheitsbl–Gesundheitsforsch–Gesundheitsschutz 49: 141–150

Knieps F (2009) Evidence Based Health Policy oder wissenschaftlich verbrämter Lobbyismus – die Verwertung wissenschaftlicher Erkenntnisse in der Gesundheitspolitik. ZEFQ 103: 273–280

Kuhn J (2007) Editorial zum Schwerpunktheft »Evaluation in der Prävention und Gesundheitsförderung«. Prävention 30: 7

Leyland AH (2010) Methodological challenges in the evaluation of community interventions. European Journal of Public Health 20: 242–243

Milton B, Moonan M, Taylor-Robinson D et al. (2011) (Eds) How can the health equitiy impact of universal policies bei evaluated? World Health Organization, Copenhagen

Pawson R, Greenhalgh T, Harvey G et al. (2005) Realistic review – a new method of systematic review designed for complex policy interventions. J Health Serv Res Policy 10 (Suppl 1): 21–34

Sackett DL, Rosenberg WMC, Gray JAM et al. (2012) Was ist EBM und was nicht? In: Moser G, Stöckel S, Kuhn J (2012) Die statistische Transformation der Erfahrung. Centaurus, Heidelberg, S 13–17

Victora CG, Habicht JP, Bryce J (2004) Evidence-Based Public Health: Moving Beyond Randomized Trails. American Journal of Public Health 94: 400–405

Teil 1
Evaluation komplexer Intervention in der Prävention, ein Problemaufriss: Lernende Systeme und ihr »Datenbild«, Modellierungsmöglichkeiten und -grenzen

Möglichkeiten und Grenzen der Evaluierbarkeit komplexer Interventionen

Hans-Peter Lorenzen

Seit 2008 veröffentlicht die DeGEval – Gesellschaft für Evaluation Positionspapiere, welche die Ergebnisse der jeweiligen Jahrestagung prägnant zusammenfassen. Unter der Überschrift »Steuerung braucht Evaluation« beginnt das Positionspapier 01 mit der Feststellung: »Moderne Gesellschaften sind geprägt durch komplexe Steuerungsprozesse, in denen Wirtschaft, Politik, Bildung, Sozialsystem und weitere Akteure und Akteurinnen vielfältige, zum Teil gegenläufige Interessen mit unterschiedlichen Handlungslogiken verfolgen« (DeGEval 2011).

Die DeGEval geht zu Recht davon aus, dass Evaluationen zu einer evidenzbasierten Politik und Praxis beitragen können. Zu allererst aber muss diese Funktion von den jeweiligen Entscheidungsträgern auch gewollt werden. Klassische Politik lebt häufig von Kompromissen, für deren Zustandekommen konkrete Evaluationsergebnisse eher hinderlich sein können. Aber auch wenn konkrete Ergebnisse gewollt werden, ist es alles andere als selbstverständlich, dass man mit Evaluationen diese umfassende Perspektive auch einlösen kann.

Vergleicht man diese Perspektive mit der Entstehungsgeschichte der DeGEval – Standards, so fällt auf, dass Standards in den USA 1981 zunächst für die Evaluation von Programmen im Bildungsbereich entwickelt worden sind. Diese wurden 2001 fast vollständig mit den Standards der SEVAL in der Schweiz und der DeGEval in Deutschland übernommen und für einen weit größeren Anwendungsbereich empfohlen (Deutsche Gesellschaft für Evaluation 2002).

Obwohl in der von der DeGEval genannten Definition einer Evaluation auch Evaluationsgegenstände wie Organisationen oder Politik genannt werden, die auf komplexe Interventionen hindeuten, findet sich in den Einzelstandards und den zugehörigen Erläuterungen keine Differenzierung nach dem Komplexitätsgrad der Interventionen. Vielleicht war daran gedacht, auch diesen Aspekt bei der zunächst für 2004 vorgesehenen Überarbeitung der Standards und Ergänzung durch ausführliche Hinweise für die praktische Anwendung Rechnung zu tragen. Diese Überarbeitung wurde bisher nicht geleistet. Stattdessen veröffentlichte die DeGEval Positionspapiere und themenbezogene Broschüren mit Empfehlungen zur Selbstevaluation, für die Aus- und Weiterbildung von Evaluatoren und für Auftraggeber von Evaluationen.

Vielleicht ist es an der Zeit, die in verschiedenen Gegenstandsbereichen gemachten Erfahrungen mit unterschiedlichen Formen von Evaluation komplexer Interventionen zu sammeln und systematisch auszuwerten. Lassen Sie mich mit dieser schwierigen Aufgabe einen Anfang machen.

Bereich 1
Förderprogramme der Forschungs-, Technologie- und Innovationspolitik

Der Arbeitskreis FTI-Politik der DeGEval hat bereits 2000 für seinen Gegenstandsbereich Potenziale der Evaluation von Multi-Akteur-/Multi-Maßnahmen-Programmen betrachtet. Bührer und Kuhlmann haben die Ergebnisse 2003 veröffentlicht (Bührer, Kuhlmann 2003). Zu der Zeit lagen umfangreiche Erfahrungen mit der Evaluation vergleichsweise einfacher Förderprogramme zur Technologieentwicklung vor, bei denen man z. B. die Anschlussfähigkeit von einer Phase der Wertschöpfungskette an die nächste als Erfolg definieren konnte. Aber auch schon hier setzten Zeitfaktor und Zurechenbarkeit einem Wirkungsnachweis enge Grenzen. Wirkungen sind meist dann nicht schon nachweisbar, wenn politisch über Fortführung oder Abbruch eines Programms entschieden werden muss. Zwar lassen sich Veröffentlichun-

gen und Patentanmeldungen noch gut zählen, aber Produktentwicklungen z. B. brauchen Jahre, sodass unmittelbar nach Abschluss einer Vorentwicklungsphase noch keine Aussagen zur Wirkung gemacht werden können. Was die Zurechenbarkeit angeht, fließt in Produktentwicklungen eine Fülle von Vorerfahrungen des betreffenden Unternehmens ein, sodass der Anteil einer geförderten Vorentwicklung schwer zu beziffern ist.

In den 1990er-Jahren setzte sich die Auffassung durch, dass Innovationsprozesse nicht linear, sondern interaktiv und mit Rückkopplungsschleifen verlaufen. Erfolgreiche Innovationen sind häufig durch disziplinübergreifende Kooperationen jeweils mehrerer Forschungseinrichtungen und Unternehmen gekennzeichnet, zu denen dann noch weitere Akteure wie Finanzdienstleister, Bildungseinrichtungen oder Normungsgremien beitragen. Der Wissensentstehung durch individuelle wissenschaftliche Neugier wurde als neues Paradigma die anwendungsgetriebene Wissensentstehung an die Seite gestellt (Kuhlmann 2002).

Diese Einsichten führten zu immer komplexeren Förderprogrammen, mit denen z. B. die Bildung regionaler Innovationsnetzwerke mit technologischen Schwerpunkten oder Technologien freier Wahl angeregt und über deren Förderung im Wettbewerb entschieden wurde. Eine Evaluation kann hier genauso wie bei den alten einfachen Programmen nach den Effekten bei den einzelnen Forschungseinrichtungen und Unternehmen fragen und darüber hinaus nach Kooperationsverhalten und Kooperationserfolg. Nur ist die Zurechenbarkeit hier noch einmal schwerer zu bestimmen, besonders dann, wenn nicht nur die spezifischen Kooperationskosten gefördert werden, sondern auch Zuschüsse zur Technologieentwicklung und zum Aufbau von Infrastruktur gewährt werden.

Anders als bei einfachen Förderprogrammen, die sich an kleine und mittlere Unternehmen richten und diesen die Auswahl ihrer Technologie überlassen, lässt sich hier kein quasi-experimentelles Evaluationsdesign mit matched pairs zugrunde legen. Die Evaluatoren können die Fragen nach Wirkungen und nach einer Kosten-/Nutzen-Relation nicht mehr beantworten, sie konzentrieren sich infolgedessen – wie in der Veröffentlichung von Bührer und Kuhlmann dargelegt wird – auf formative, lernorientierte Aspekte. Sie unterstützen die staatliche Seite bei ihren Aushandlungsprozessen durch Aufbereitung von Akteursperspektiven und helfen mit, verfestigte Akteursorientierungen zu überwinden (Bührer, Kuhlmann 2003).

Bereich 2
Forschungs- und Forschungsförderungseinrichtungen im Wandel

Universitäten und außeruniversitäre, öffentlich geförderte FuE-Einrichtungen werden inzwischen regelmäßig evaluiert. Je nach Verortung im Innovationssystem werden die wissenschaftliche Qualität oder der Transfer der FuE-Ergebnisse in die Wirtschaft oder deren Verwendung durch die Öffentliche Hand geprüft. Das beantwortet aber nur den zweiten Teil der Frage: Machen wir das Richtige richtig?, nämlich, ob die durchgeführten Arbeiten richtig gemacht wurden. Ob die Themen der FuE-Arbeiten und die dafür beschaffte Ausrüstung richtig gewählt worden waren, bleibt offen. Die Gefahr besteht darin, dass neue, meist interdisziplinäre und von der Anwendung getriebene Felder vernachlässigt werden, die nicht in die bisherige Organisation passen. Die Kernfrage einer auf die Zukunft gerichteten Evaluation heißt dann: Besitzt die Einrichtung geeignete Verfahren, mit denen solche Verkrustungen erkannt und abgebaut, sowie neue zukunftsgerichtete Felder identifiziert und implementiert werden können (Weule 2002). Hier wird Evaluation zur Managementberatung. Die Antworten müssen die Einrichtungen ausgehend von ihren Stärken selbst finden. Die Vorstellung, staatliche Stellen könnten bevorzugte Themen, z. B. aus einem Delphi-Prozess vorgeben, gehen an der Realität vorbei. Eher können die Einrichtungen solche Delphi-Themen als Check-Liste in ihren internen Diskussionen nutzen. Die Prioritätensetzung ist auch bei FuE-Einrichtungen eine unternehmerische Entscheidung, genauso wie wenn in einem Unternehmen die Bearbeitung neuer Märkte beschlossen wird.

Das wirft noch deutlicher als im vorangegangenen Beispiel ein neues Licht auf das Verständnis von Steuerung, dem Begriff, mit dem ich meinen Vortrag begonnen habe. Staatliche Steuerung bedeutet dann, nach Kuhlmann "a government-led 'mediation' between diverging and competing interests of various players within the science and technology system" (Kuhlmann 2002).

Diese Einsicht gilt aber nicht nur für das System von Wissenschaft und Technologie. Eine entsprechende Vermittlung beabsichtigt die »Roadmap für das Gesundheitsforschungsprogramm der Bundesregierung« von 2007 (Gesundheitsforschungsrat des BMBF 2007). Sie soll eine »Orientierung auf dem Weg zu den zukünftig wichtigen Themen der Gesundheitsforschung bieten«. Sie konzentriert sich auf die großen Volkskrankheiten und steht damit in der Tradition der Gesundheitsforschungsprogramme der Bundesregierung seit 1978 (Bundesminister für Forschung und Technologie 1978). Neu ist die Forderung nach engerer Verschränkung von klinischer Forschung und Grundlagenforschung. Diese Verschränkung entspricht dem erwähnten neuen Paradigma von Wissensentstehung im Bereich der Technologie, dass nämlich Themen der Grundlagenforschung von den brennenden Fragen der Klinik angeregt werden sollten. Der Gesundheitsforschungsrat des Bundesministeriums für Bildung und Forschung hofft, »dass andere Einrichtungen der Gesundheitsforschung, Fördereinrichtungen, Forschungsinstitute und auch medizinische Fakultäten in der Ausrichtung ihrer Förder- und Forschungsstrategien Orientierungen und Anregungen in der Roadmap finden werden.« Inhaltlich eröffnet die Roadmap als Exante-Evaluation neue Perspektiven für Entscheidungsprozesse, ohne Ziele im Einzelnen vorzugeben, und vermittelt damit zwischen den divergierenden Interessen der Akteure im Gesundheitssystem.

Bereich 3
Öffentlich zugängliche Daten als Basis für Entscheidungen

Hierzu gehört die »Landkarte Hochschulmedizin«, die z. Z. in dritter Auflage vorliegt (Fraunhofer Institut für System- und Innovationsforschung 2007). Sie stellt im Internet Daten zu Lehre, Forschung und Krankenversorgung, sowie zu Personal, Finanzen und Struktur zur Verfügung und schafft so Leistungstransparenz. Sie »verzichtet auf einseitige Wertungen und Rankings, eröffnet (aber) vielfältige Nutzungsmöglichkeiten«, und zwar für sehr unterschiedliche Zielgruppen, wie z. B. Medizinische Fakultäten, Universitätskliniken, Wissenschaftsadministration, Wissenschaftler, Studenten und Wirtschaftsvertreter. Auch in diesem Fall wurden mit einer Evaluation neue Diskussionen angestoßen.

Geleistet wurde diese Arbeit in der Verantwortung des Bundesministeriums für Bildung und Forschung und des Medizinischen Fakultätentages. Wenn man die Aufgabenteilung von Bund und Ländern bedenkt, spricht viel dafür, dass die eigentlich zuständigen Akteure sich dieser Aufgabe selbst nicht stellen wollten. Das würde auch erklären, weshalb es, wie Kuhlmann beklagt, eine »bundesweite Übersicht der universitären Fachdisziplinen bisher nicht gibt« (Kuhlmann 2009). Umso erfreulicher ist, dass die anderen Akteure den Mut gehabt haben, das heiße Eisen anzufassen, und dass Wettbewerb sehr wohl auch im Rahmen staatlichen Handelns vorkommt.

Ganz andere Grenzen werden sichtbar bei Monitoring und Evaluation von Stadt- und Regionalentwicklung, die wesentlich durch das Gebot der Nachhaltigkeit geprägt werden. Monitoring bedeutet hier die Überwachung der Umweltauswirkungen. Die Evaluation zielt »auf die Verfolgung der beabsichtigten Auswirkungen von Plänen und Programmen, d. h. im Wesentlichen auf die Beurteilung ihrer Steuerungsfunktion bzw. eine diesbezügliche Zielerreichung« (Jacoby 2009). Zielerreichung kann man natürlich nur messen, wenn man zusammen mit den Zielen auch Indikatoren festgelegt hat.

Solche Indikatoren findet man unter den Geodaten. Es gibt inzwischen eine Fülle von Geodaten. Ein Beispiel ist die Initiative der EU-Kommission INSPIRE (Infrastructure for Spatial Information in the European Community). Sie »verpflichtet die Mitgliedstaaten stufenweise interoperable Geobasisdaten ... sowie bereits vorhandene Geofachdaten ... bereitzustellen.« Das Problem besteht jetzt darin, »dass die meisten potenziellen Nachfrager wie Kommunen und die Bürgerinnen und Bürger damit überfordert sein dürften, das Informationsangebot und den Informationsgehalt ziel- und sachgemäß zu nutzen«. Jacoby fordert Geoinformationssysteme mit der »Einstellung von Zusatzinformationen ..., (mit denen) sich der Datennutzer direkt über die fachlich korrekte und rechtlich zulässige Datenverwendung informieren kann« (Jacoby 2009). Das würde primär den Kommunen helfen, denn aufgrund ihrer Planungshoheit sind sie Dreh- und Angelpunkt

einer nachhaltigen Entwicklung. Letztlich geht es um Verhaltensänderungen bei den Entscheidungsträgern selbst und im Verhältnis zu den Bürgerinnen und Bürgern.

Bereich 4
Paradigmenwechsel in der Gesundheitspolitik

»Die Deutsche Herz-Kreislauf-Präventionsstudie« ist wohl die komplexeste quasi-experimentelle Intervention, die je in Deutschland stattgefunden hat (Forschungsverbund DHP 1998). Sie dauerte von 1978 bis 1994, was allein schon bemerkenswert ist, weil die Finanzierung über mehrere Legislaturperioden und gegen den z. T. heftigen Widerstand der verfassten Ärzteschaft gesichert werden musste. Die Ärzteschaft fürchtete einen Paradigmenwechsel zu Lasten der kurativen Medizin. Die Studie wurde von den Antragstellern begründet durch die damals hohe kardiovaskuläre Mortalität mit Herzinfarkt und Schlaganfall in Deutschland. Sie konnten auf vorlaufende gemeindebezogene Studien in Finnland und den USA hinweisen, wo über Verhaltensänderungen in der Bevölkerung Risikofaktoren wie Zigarettenrauchen, erhöhte Cholesterinwerte, erhöhter Blutdruck und Übergewicht gesenkt werden sollten. Die Evaluation wurde sehr sorgfältig zusammen mit der Intervention geplant: Die Dauer der Intervention in der Hauptphase von 1984 bis 1991 beruhte auf der Annahme einer Latenzzeit von mindestens 7 Jahren, bevor Verhaltensänderungen zu messbaren Veränderungen der genannten Risikofaktoren führen können. Als Referenzpopulation zu den Interventionsregionen wurde die gesamte damalige Bundesrepublik, aber ohne die Interventionsregionen gewählt. Die Entwicklung der Risikofaktoren wurde in 3 Gesundheitssurveys ermittelt: Mit Ausnahme des Zigarettenrauchens nahmen in den alten Bundesländern zwischen 1984 und 1991 die Risikofaktoren zu. In den Interventionsregionen ließ sich der Trend bis auf den Risikofaktor Übergewicht aufhalten, so dass man sagen kann, dass sich das Konzept der gemeindebezogenen Intervention bewährt hat, und zwar in zweifacher Hinsicht, zum einen in Bezug auf die Veränderung des Gesundheitsbewusstseins und -verhaltens und zum anderen hinsichtlich der dafür von vielen Akteuren getragenen Infrastruktur. Nun hätte man erwartet, dass auch die kardiovaskuläre Mortalität, die anhand der Todesursachenstatistik ermittelt wurde, diese Veränderungen widerspiegelt. Das verblüffende Ergebnis war, dass in diesem Zeitraum die kardiovaskuläre Mortalität in den alten Bundesländern insgesamt zurückgegangen war, obwohl die Risikofaktoren angestiegen sind, und die kardiovaskuläre Mortalität in den Interventionsregionen noch etwas über dem Durchschnittswert lag. Die vorzügliche Dokumentation der »Deutschen Herz-Kreislauf-Präventionsstudie« verfällt am Ende in tiefe Ratlosigkeit. Die Studie scheiterte an der Menge nicht kontrollierbarer Faktoren.

Aber dabei konnte es natürlich nicht bleiben. Das Herzinfarktregister Augsburg weist bis 2007/2008 noch einmal eine gegenüber den 1980er-Jahren drastisch gesunkene Herzinfarktmortalität aus. Es gab einen weiteren Bundes-Gesundheitssurvey 1998. Zurzeit läuft die DEGS-Studie zur Gesundheit Erwachsener in Deutschland, die vom Robert Koch-Institut betreut wird. Vielleicht haben wir in der Diskussion Gelegenheit, etwas über den aktuellen Stand des Verhältnisses von Risikofaktoren und kardiovaskulärer Mortalität in ganz Deutschland zu hören.

Bereich 5
Politik als lernendes System

Mit der Hartz-Evaluation wurde ein neues Kapitel in der Wechselwirkung zwischen Politik und Politikberatung aufgeschlagen. Der Deutsche Bundestag beauftragte 2002 im Zuge des Gesetzgebungsverfahrens zum »Ersten und Zweiten Gesetz für moderne Dienstleistungen am Arbeitsmarkt« die Bundesregierung, die gesamte Reform, abkürzend nach dem Vorsitzenden der Beratungskommission »Hartz I – IV« genannt, zu evaluieren. Erste belastbare Ergebnisse wurden vor der geplanten Bundestagswahl 2006 erwartet, damit sie entweder die Reformelemente bestätigen oder ab 2007 zur Korrektur in einen weiteren Gesetzgebungsprozess einfließen könnten: Politik als lernendes System! Es ging – wie Heyer schrieb – um die Wirkungen »einerseits der Instrumente der aktiven Arbeitsmarktpolitik und der Veränderung der beschäftigungspolitischen Rahmenbedingungen und andererseits eine organisationssoziologische Analyse des Umbaus der ›alten‹ Arbeitsverwaltung zu

einem modernen Dienstleister am Arbeitsmarkt. … Weiter soll ein Bild darüber gewonnen werden, ob bzw. wie die Bürgerinnen und Bürger sowie die Kunden der damaligen Bundesanstalt für Arbeit … die Umsetzung der politischen Impulse wahrnehmen« (Heyer 2006).

Es wurde aber sehr schnell deutlich, dass eine solche Evaluation in drei Jahren nicht zu leisten war. Da war die Verzögerung bis 2005 bei der Verabschiedung von »Hartz IV«, das der Grundsicherung von Arbeitssuchenden gilt, so dass dieser Teil der Reform von der Evaluation ausgeklammert werden musste. Es fehlte eine vollständige und konsistente Baseline als Bestandsaufnahme für alle Aspekte der Reform, bevor diese wirksam werden sollten. Es mussten erst die Voraussetzungen für eine zeitnahe Versorgung der Evaluationsteams mit validen Daten geschaffen werden. Für die Evaluierung der arbeitsmarktpolitischen Instrumente und der Verbesserung der beschäftigungspolitischen Rahmenbedingungen war zusätzlich erforderlich, einen zentralen Erfolgsmaßstab zu definieren. Als einzige Lösung unter den vielen bis dahin in Deutschland üblichen sozialpolitischen Zielen wurde – entsprechend internationalen Vorbildern – die »Eingliederung in Erwerbstätigkeit« festgelegt. Dazu kamen die Anlaufzeiten der neuen arbeitsmarktpolitischen Maßnahmen (Heyer 2006). Dennoch zeigte – lt. Bericht der Bundesregierung – die Evaluation im Ganzen, dass die Reform in die richtige Richtung ging. Einzelne Instrumente wurden noch 2006 modifiziert, so z. B. der Existenzgründungszuschuss (Ich-AG). Neue Voraussetzungen waren dann persönliche und fachliche Eignung des Gründers sowie ein tragfähiger Geschäftsplan (Deutscher Bundestag – Wissenschaftliche Dienste – 2006).

Fazit

Kann man nun aus den verschiedenen Praxisbeispielen zu Möglichkeiten und Grenzen der Evaluation komplexer Interventionen verallgemeinerbare Aspekte ableiten, die Grundlage von Empfehlungen sein können? Die Vielfalt der gerade berichteten einzelnen Befunde unterstreicht meine These, dass wir mit diesem Prozess erst am Anfang stehen. Ohne eine breite Bestandsaufnahme und anschließende Analyse geht es nicht!

Literatur

Bührer S, Kuhlmann S (Hrsg) (2003) Politische Steuerung von Innovationssystemen? Fraunhofer IRB Verlag, Stuttgart

Bundesminister für Forschung und Technologie (1978) Programm der Bundesregierung zur Förderung von Forschung und Entwicklung im Dienste der Gesundheit 1978–1981, Bonn

DeGEval-Gesellschaft für Evaluation (2011) Publikationen/Positionspapiere www.degeval.org/publikationen/positionspapiere (Stand: 28.07.2011)

Deutsche Gesellschaft für Evaluation (2002) Standards für Evaluation, Köln

Deutscher Bundestag – Wissenschaftliche Dienste – (2006) Die Evaluation von Hartz I – III im Überblick – Info-Brief – www.bundestag.de/dokumente/analysen/2007/Die_Evaluierung_von_Hartz_I-III_im_Ueberblick.pdf

Forschungsverbund DHP (Hrsg) (1998) Die Deutsche Herz-Kreislauf-Präventionsstudie. Verlag Hans Huber, Bern, Göttingen, Toronto, Seattle

Fraunhofer Institut für System- und Innovationsforschung (ISI) (2007) Landkarte Hochschulmedizin, Karlsruhe www.landkarte-hochschulmedizin.de/landmapmain.aspx

Gesundheitsforschungsrat des BMBF (2007) Roadmap für das Gesundheitsforschungsprogramm der Bundesregierung, Bonn

Heyer G (2006) Zielsetzung und Struktur der »Hartz-Evaluation«. Zeitschrift für Arbeitsmarktforschung 39 (3/4): 467–476

Jacoby C (2009) Monitoring und Evaluation von Stadt- und Regionalentwicklung. Einführung in Begriffswelt, rechtliche Anforderungen, fachliche Herausforderungen und ausgewählte Ansätze. In: Jacoby C (Hrsg) Monitoring und Evaluation von Stadt- und Regionalentwicklung. Akademie für Raumforschung und Landesplanung, Hannover, S 1–24

Kuhlmann S (2002) Governance and Intelligence in Research and Innovation Systems, Inaugural Lecture. Universiteit Utrecht, Utrecht

Kuhlmann S (2009) Evaluation von Forschungs- und Innovationspolitik in Deutschland – Stand und Perspektiven. In: Widmer T, Beywl W, Fabian C (Hrsg) Evaluation. Ein Systematisches Lehrbuch. VS Verlag für Sozialwissenschaften, Wiesbaden, S 283–294

Weule H (2002) Abschlussbericht der Evaluationskommission. Evaluation der Physikalisch-Technischen Bundesanstalt. Bundesministerium für Wirtschaft und Arbeit, Berlin

Vor dem Messen und Rechnen: Die Landschaft beschreiben

Überlegungen für eine Klassifizierung und einheitliche Terminologie von Gesundheitsförderungsinterventionen als Voraussetzung für Evaluation und Evidenzbildung

Alf Trojan

Das Ziel, die Evaluation komplexer Gesundheitsförderungsinterventionen zu verbessern, dient einem weiter reichenden Ziel, nämlich eine verlässliche Evidenzbeurteilung der Wirksamkeit bestimmter Maßnahmen zu ermöglichen. Ein 2011 erschienener Bericht zur Evaluation von Maßnahmen für mehr gesundheitliche Chancengleichheit (Milton et al. 2011) stellt den Mangel an Evidenz als Haupthinderungsgrund für mehr Handeln auf Bevölkerungsebene heraus. Obwohl der Bericht sich primär auf die nationale und internationale Ebene bezieht, können wir auf allen Ebenen sowohl bezüglich der Problemanalyse wie auch der viel versprechenden Ansätze daraus lernen. Der Bericht macht auch mit einigen knappen Beispielen deutlich, dass die Umsetzung allgemeiner, auf ganze Bevölkerungen gerichtete Maßnahmen (universal policies) auf der lokalen Ebene besonders leicht fehlschlagen kann (S. 6). Der vorliegende Beitrag ist von seinem empirischen Hintergrund her stark auf diese Ebene abgestellt, versteht sich aber als grundsätzlicher Beitrag zur Frage, wie wir die Evidenzbeurteilung von komplexen Interventionen der Gesundheitsförderung und Prävention verbessern können.

Im *ersten* Abschnitt wird die besondere Komplexität des diesem Beitrag zugrunde liegenden Praxisprojekts aufgezeigt und die Fragestellung entwickelt. Im *zweiten* Abschnitt möchte ich begründen, warum ich die in der Überschrift genannte Maxime für nur scheinbar trivial halte und einen Blick auf häufig gebrauchte allgemeine Oberbegriffe werfen. Im *dritten* Abschnitt geht es um die Charakterisierung komplexer Interventionsprogramme in der Gesundheitsförderung und den Status Quo des allgemeinen Sprachgebrauchs. Im *vierten* Abschnitt werden erfahrungs-gegründete Vorschläge gemacht für die Strukturierung und Klassifizierung von weniger komplexen Interventionen (Interventionsbausteinen) in Spezialprogrammen sozialraumbezogener Gesundheitsförderung. Im abschließenden Abschnitt wird unterbreitet, wie der Evidenzaufbau für Interventionen der Gesundheitsförderung organisiert werden könnte.

1 Verortung der quartiersbezogenen Gesundheitsförderung in der Landschaft komplexer Gesundheitsförderungsprogramme

Auf der Tagung, aus deren Kontext dieser Beitrag hervorgeht, war nicht die ganze Breite, sondern nur ein Ausschnitt komplexer Interventionsprogramme der Gesundheitsförderung vertreten. Dieser Ausschnitt betraf vor allem landes- und bundesweite Dachprogramme für Gesundheitsförderung und Prävention, so dass man den Eindruck hätte gewinnen können, dass Komplexität vor allem dadurch entsteht, dass sich die Programme auf höheren politischen Ebenen als der lokalen abspielen. Auch in einem anderen Punkt gab es eine Gemeinsamkeit der vorgestellten Präventionsprogramme, die nicht immer und überall typisch ist für komplexe Interventionen: Bei dem nationalen Aktionsplan »IN FORM« und bei der bayrischen Gesundheitsinitiative »Gesund. Leben.Bayern.«, aber auch bei den anderen Programmen, ging es um breit und in der Regel langjährig oder sogar auf Dauer angelegte *Förderstrategien* für Interventionsprogramme. Vernetzungs- und Koordinierungsaktivitäten (Capacity Building) spielen in allen Programmen eine große Rolle, egal ob es sich um (nach Art von Projekten) zeitlich begrenzte oder um Regelaufgaben handelt (Prävention nach § 20 SGB V, Deutsche Arbeitsschutzstrategie, Fonds Gesundes Österreich). Ein herausragendes gemeinsames Merkmal ist bei diesen Programmen, dass erhebliche finanzielle Mittel eingesetzt werden für Interventionen und dass bei der Evaluation natürlich primär die Frage im Raum steht, ob diese Mittel sinn- und wirkungsvoll eingesetzt werden.

Im Tagungsprogramm waren *nicht* explizit vertreten die deutschen und internationalen Settingnetzwerke oder landesweite Koordinierungsprogramme, wie z. B. der Pakt für Prävention in Hamburg. Charakteristisch für diese aber auch viele andere Gesundheitsförderungsaktivitäten auf lokaler Ebene ist, dass es zwar Akteure und Koordinie-

rungsstellen gibt, dass aber kein großer Topf für die Finanzierung von Interventionen vorhanden ist. Dieser Unterschied hat erhebliche Bedeutung für die Planung, Durchführung und Evaluationsmöglichkeiten solcher Programme.

Der Programmkontext, aus dem mein Beitrag entstanden ist, unterscheidet sich also von den anderen auch in diesem Buch vertretenen Programmen zumindest in drei Aspekten:

- er gehört zur Programm-Familie der settingbezogenen Gesundheitsförderung,
- es handelt sich um die niedrigste sozialräumliche Programmebene, nämlich die Gesundheitsförderung im Quartier (vgl. allgemein zu Ansätzen auf dieser Ebene Trojan, Süß 2011),
- es gab fast kein Geld für Interventionen (alles musste aus den Regelaufgaben der Beteiligten mit geringsten Mitteln geleistet werden) und ausnahmsweise durch das Forschungsprogramm vergleichsweise sehr viel Geld für die begleitende Evaluation.

Die hier beispielgebende quartiersbezogene Gesundheitsförderung, die nach dem Namen des Viertels (Lenzsiedlung) »Lenzgesund« genannt wurde, hat zwei entscheidende Komplexitätsfaktoren, nämlich erstens den *doppelten* Interventionsfokus, der sowohl die *Quartiersentwicklung* wie auch die *Gesundheitsförderung* im Quartier ins Visier nimmt. Damit ist es auch ein »echter« settingbezogener Ansatz, der das Setting nicht nur als Zugang zu selektiven Zielgruppen benutzt. Der zweite Komplexitätsfaktor liegt darin, dass es eigentlich drei voneinander unterscheidbare (wenn auch eng verwobene) Interventionsebenen gibt: das Quartier als Ganzes, das definierte »Präventionsprogramm Lenzgesund« mit acht Handlungsbereichen rund um Schwangerschaft, Geburt und die ersten Lebensjahre sowie zahlreiche Einzelinterventionen, die in diesem Rahmen für die entsprechenden Zielgruppen organisiert wurden (vgl. ausführlicher Mossakowski et. al. 2010).

Nimmt man die Frage aus dem Titel dieses Buches auf, lässt sich schon an dieser Stelle und als Charakterisierung der quartiersbezogenen Gesundheitsförderung im Lenzviertel sagen: Der Ausdruck des »lernenden Systems« ist außerordentlich passend, weil das Programm sehr stark die Bürger zu beteiligen versucht hat (Runder Tisch, Bürgerbefragungen, Evaluation durch Rückmeldungen von Interventionsteilnehmer/-innen etc.) und weil die Rückmeldungen, großenteils vermittelt über die Begleitforschung, direkt in die Gestaltung und Anpassung des Programms eingeflossen sind. Ein solches dynamisches lernendes System mit den berühmten “moving targets”, also den sich bewegenden Zielen, stellt die Evaluation vor besondere Herausforderungen.

Ganz sicher ist das Programm auch als »lehrreiches System« zu bezeichnen. Einmal hängt dies schlicht damit zusammen, dass viel empirisch fundiertes Wissen gesammelt und in diesem Fall durch die großzügige Begleitforschung auch gut dokumentiert wurde. Ein anderer wichtiger Aspekt ist, dass das Programm inzwischen seit über 10 Jahren läuft, das heißt, es gab auch genügend Zeit, die Lehren des Programms aufzunehmen und im weiteren Verlauf umzusetzen.

All dies Gesagte bezieht sich primär auf die Quartiersebene und die beteiligten Akteure. Durch die Begleitforschung, aber auch andere Medien wie die Projektdatei des Kooperationsverbundes »Gesundheitliche Chancengleichheit« sind unseres Erachtens auch viele Erfahrungen über das von uns beforschte Quartier hinaus bekannt gemacht worden, so dass exemplarisches Lernen an diesem Modell vermutlich sehr viel besser ermöglicht wurde, als es sonst bei lokalen Projekten üblich ist.

Mehr oder weniger explizit formuliert hat das Projekt jedoch weiter reichende Ziele: Eigentlich wird gewünscht, dass es eine möglichst universell gültige Aussage macht, inwieweit der allgemeine Typus »quartiersbezogene Gesundheitsförderung« funktioniert, dass heißt, ob dieser Interventionstypus allgemein als wirksam empfohlen werden kann. Um wirklich als methodisch genügend belastbarer Baustein für Aussagen zur Evidenzlage bei sozialraumbezogener Gesundheitsförderung zu dienen, sind verschiedene Hürden zu nehmen, die dieses Ziel fast utopisch erscheinen lassen. Eine der wichtigsten Voraussetzungen wäre die Integration der verschiedenen Teilevaluationen, die für die verschiedenen Programmebenen (siehe vorn) durchgeführt worden sind. Eine weitere, für diesen Beitrag zentrale Voraussetzung ist, dass sowohl auf der obersten Programmebene, wie auch auf den unteren eindeutig definiert und benannt werden kann, auf welche Intervention sich der jeweilige Evaluationsansatz bezieht.

Als wir uns im Projekt diese Frage klar gemacht hatten, begannen unsere Überlegungen, wie denn das Programm als Ganzes eindeutig zu bezeichnen wäre und wie seine Bausteine im Einzelnen aussehen und benannt werden können.

2 »Kartierung« der Landschaft von Oberbegriffen und Ansätzen im Bereich komplexer Gesundheitsförderungsinterventionen

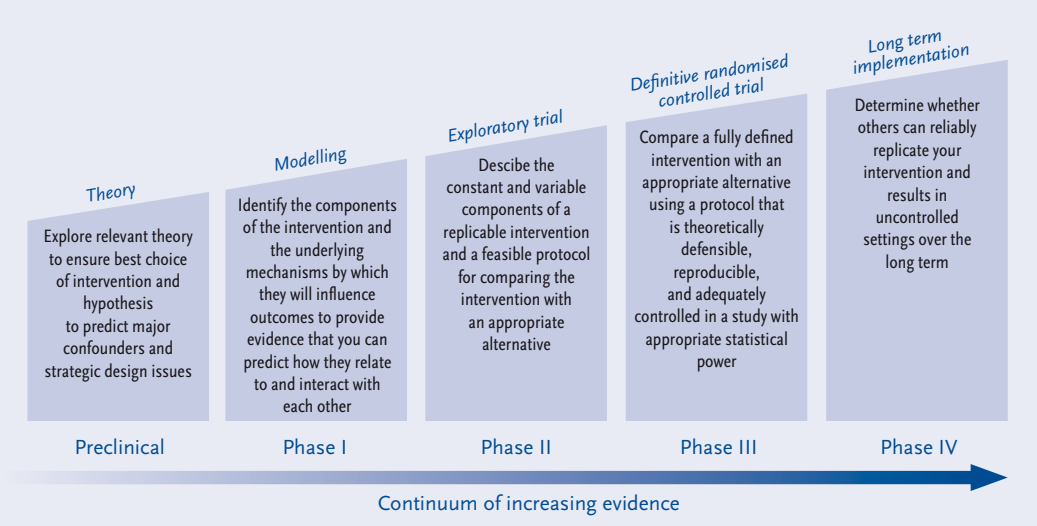
Die Forderung, die Landschaft von Interventionen der Gesundheitsförderung genau zu beschreiben bzw. zu definieren, erscheint auf den ersten Blick so selbstverständlich, dass sie fast trivial wirkt:

- Man braucht allgemein übereinstimmend definierte Ausdrücke für die Verständigung in der Wissenschaft, aber auch für die Verständigung zwischen den Bereichen Praxis, Forschung und Politik.
- Erst durch die differenzierte Betrachtung und Benennung einzelner Bestandteile eines komplexen Interventionsprogramms wird deutlich, wie es sich von gleichartig und ähnlich benannten unterscheidet. Dies ist für die Übertragbarkeit von Erfahrungen von großer Bedeutung.
- Aktuell sind an verschiedenen Stellen im Internet Verzeichnisse von Evaluationsinstrumenten für die Nutzung durch Forschung und Praxis abgelegt worden (z. B. www.evaluationstools.de; Töppich, Linden 2011). Für diese Evaluationsinstrumente müssen stringenter als bisher die Arten/Typen von Interventionen angegeben werden, für die das jeweilige Instrument geeignet ist.
- Und schließlich sind fast in jeder Phase der Entstehung von Evidenz eindeutige Angaben darüber nötig, um welche Komponenten von Interventionen es sich in der jeweiligen Phase handelt (vgl. Campbell 2000, Abbildung 1).

Es gibt also genügend gute Gründe, zunächst einmal die Frage zu klären, wie viel Einigkeit und Einheitlichkeit bei der Verwendung der wichtigsten Oberbegriffe für Präventions- und/oder Gesundheitsförderungsprogramme es gibt. Dabei sind wir so vorgegangen, dass wir die gesundheitspolitisch wichtigsten Akteursgruppen bzw. wichtige Referenzsysteme und ihre Terminologie angeschaut haben.

Qualität in der Prävention (QIP-System) – In diesem System wird vor allem der Begriff Projekt definiert als »ein abgegrenztes, in sich durchdachtes und zusammenhängend aufgebautes Vorhaben

Abbildung 1
Sequenzielle Phasen der Entwicklung und Evaluation komplexer Interventionen
Quelle: Campbell M et al. BMJ 2000, 321: 694–696 (reproduziert mit Genehmigung der BMJ Publishing Group, 8/2012)



zur Prävention und Gesundheitsförderung«. Ein Projekt könne man erkennen an:

- selbstständiger Zielsetzung, Konzeption und Planung,
- gesondertem Durchführungsauftrag,
- eigener Ausstattung oder Budget,
- besonderem Namen,
- personell oder organisatorisch geregelter Zuständigkeit.

Es wird nicht gesagt, dass alle diese Kriterien notwendigerweise da sein müssen, sondern nur »zum Beispiel«. Es wird auch nicht angesprochen, dass der Projekt-Begriff eng verknüpft ist mit zeitlich begrenzter Laufzeit und genau aus diesem Grunde in der Gesundheitsförderung sogar mit negativen Assoziationen belegt ist (»Projektitis« statt Gesundheitsförderung als Regelaufgabe).

Gesetzliche Krankenversicherung (GKV gemäß §20, 1 und 2 im SGB V) – In den gemeinsamen und einheitlichen Handlungsfeldern und Kriterien der Spitzenverbände der Krankenkassen zur Umsetzung von §20 wird unterschieden zwischen 1. Settingansatz, 2. individueller Prävention und 3. betrieblicher Gesundheitsförderung. Als Grobeinteilung ist dies durchaus hilfreich, wenn auch von der Systematik her nicht ganz befriedigend, da ja auch die betriebliche Gesundheitsförderung einen Settingansatz beinhaltet.

Bundeszentrale für gesundheitliche Aufklärung und Kooperationsverbund Gesundheitliche Chancengleichheit – Hier spielt der Ausdruck »Programm« eine Rolle als »konzeptioneller und organisatorischer Rahmen für Projekte«. Unter »Projekt« wird verstanden die »Arbeit mit Zielgruppen«. Auch der Ausdruck »Netzwerkstrukturen« spielt eine große Rolle für »Koordination und Vernetzung«. Weiterhin gibt es umfangreiches Material über den »Good Practice Prozess« und in der Datei der Gesundheitsprojekte tauchen natürlich alle Begriffe auf, die überhaupt eine Rolle in diesem Kontext spielen. Im Glossar des Good Practice Prozesses finden sich folgende durchaus akzeptablen, aber auch sehr allgemeinen Definitionen:

> »Eine **Maßnahme** ist eine konkrete Handlung oder ein Set konkreter Handlungen mit festgelegten Terminen und Verantwortlichkeiten, die ergriffen wird, um ein Ziel oder Zwischenziel zu erreichen. Im Gegensatz zu Projekten sind hier in der Regel zeitlich unbefristete Regelangebote gemeint.«

> »**Projekte** sind komplexe, räumlich, zeitlich und finanziell begrenzte Vorhaben, die auf bestimmte Ziele ausgerichtet sind und bestimmte Resultate hervorbringen sollen« (Gesundheit BB e. V. und BZgA 2012).

Bundesvereinigung für Prävention und Gesundheitsförderung (BVPG) – 2010 hatten sich die Mitgliedsorganisationen der Bundesvereinigung Prävention und Gesundheitsförderung e. V. (BVPG) über ihre eigenen Aktivitäten in der Qualitätsentwicklung ausgetauscht. Drei Aufgaben gaben sie dem Vorstand und der Geschäftsstelle damals mit auf den Weg, darunter als erste auch, »einen einheitlichen Sprachgebrauch für die Qualitätsentwicklung in Prävention und Gesundheitsförderung zu entwickeln« (BVPG 2011). In dem hier zitierten Statusbericht 4 werden alle denkbaren Oberbegriffe verwendet. Dabei wird auch deutlich, dass sogar bezüglich Qualitätsentwicklung bei der BZgA ein konkurrierender Begriff verwendet wird, nämlich »Qualitätsverbesserung«. Ein Bericht, in dem zentrale Begriffe vereinheitlicht definiert werden, wird für 2012 angekündigt.

Die Hoffnung, dass sich aus dem Sprachgebrauch der wichtigsten Akteursgruppen eine einheitliche Terminologie oder sogar eine Klassifikation, Typologie oder ähnliches herausdestillieren lässt, wird also enttäuscht. Die Frage ist, ob nicht auch diese Akteure ein Interesse daran zeigen müssten, für die Gesundheitsförderung und Prävention ein möglichst stringent gegliedertes, hierarchisches System verschiedener Ebenen zu haben, mit denen die Landschaft der Gesundheitsförderung und Prävention beschrieben werden kann. Dies wäre die Voraussetzung, um systematisch geordnet Evidenzaussagen zu den verschiedenen, mehr oder weniger komplexen Interventionen in der Gesundheitsförderung und Prävention sammeln und zur Verfügung stellen zu können.

3 Komplexe Interventionen der Gesundheitsförderung: gängige Bezeichnungen und ihre Definition

Im Rahmen der Verortung unseres eigenen Ansatzes der quartiersbezogenen Gesundheitsförderung haben wir aus der einschlägigen Literatur und dem Sprachgebrauch einige komplexe Interventionen festgehalten, sie in möglichst hohem Einklang mit dem allgemeinen und wissenschaftlichen Sprach-

gebrauch definiert und versucht, sie anhand einiger gängiger Kriterien voneinander zu unterscheiden (Für eine aktuelle Übersicht über den Diskussionsstand zu komplexen Interventionen verweise ich auf ein Referat Martin Härters, 2011, das er mir freundlicherweise zur Verfügung gestellt hatte).

Der oberste und allgemeinste Begriff dafür, um was es sich handelt, ist »Interventionen«, also Eingriffe in ein sonst anders ablaufendes Geschehen. Dabei muss als Erstes unterschieden werden zwischen Maßnahmen der Gesundheitsförderung, die fest institutionalisiert sind und solchen, die als Projekt oder Programm mit unterschiedlichen Laufzeiten begrenzt angelegt sind.

Gesundheitsförderung als *Regelaufgabe* (also z. B. gemäß § 20 SGB V) ist gegründet auf Gesetz, Verordnung, Beschluss oder ähnliches und ist vom Prinzip her zeitlich unbegrenzt bezüglich des Auftrags (obwohl natürlich ein Auftrag durch Gesetzesänderung oder ähnliches zurückgenommen oder verändert werden kann). Im Prinzip ist die Regelaufgabe auch inhaltlich unbestimmt, das heißt der Auftrag erstreckt sich auf das Ganze der Gesundheitsförderung und Prävention. Da aber die Ressourcen zwar im Prinzip dauerhaft verfügbar, in der Höhe aber begrenzt sind, wird die Regelaufgabe häufig auf Schwerpunkte, Prioritäten, einzelne Ziele oder ähnliches eingegrenzt. Auf diese Weise hat die konkrete Durchführung von Regelaufgaben häufig auch Projekt- oder Programmcharakter. Regelaufgaben der Gesundheitsförderung und Prävention gibt es auf allen politischen Ebenen: Bundesebene (GKV), Landesebene (ÖGD-Gesetze), Bezirk (offizielle Beschlüsse der Bezirksentwicklungsplanung), Betriebe (Arbeitssicherheitsgesetz und andere Vorschriften). Das Präventionsprogramm Lenzgesund des Gesundheitsamtes Eimsbüttel hatte von Anbeginn den Charakter eines langfristig angelegten zunächst »Projektes«, dann »Programms«, und wurde später als eines der »Leuchtturm-Projekte« der Bezirksentwicklungsplanung bezeichnet. Auch viele Aktivitäten in der Lenzsiedlung, z. B. des Jugendamtes, des Sportclubs, der Schulen hatten ihre Begründung in Regelaufgaben.

Zur Definition eines *Programms* wird oft der Ausdruck eines »Bauplans« für aufeinander abgestimmte Maßnahmen mit relativ festgelegten erprobten »Bausteinen« (z. B. auf Basis von Manualen, Handbüchern, Veranstaltungsformen) benutzt. Das Programm ist in aller Regel zeitlich begrenzt, inhaltlich auf einzelne Handlungsbereiche festgelegt und kann Ziele völlig unterschiedlicher Reichweite habe (vom einfachen Raucherentwöhnungsprogramm bis hin zu einem Strategieprogramm wie der deutschen Arbeitsschutzstrategie). Der Begriff des Programms ist völlig offen für verschiedene Inhalte. Eine zentrale Dimension sind Festgelegtheit bzw. Vorgegebenheit und partizipative Weiterentwicklungsmöglichkeiten von Programmen. Normativ fordert eigentlich die Gesundheitsförderung für jedes Programm die Partizipation der Betroffenen und damit auch Veränderungsmöglichkeiten entsprechend ihren Rückmeldungen. In der Realität ist dies allerdings längst nicht immer der Fall. Uns erscheint es wichtig und sinnvoll, bei der Charakterisierung von Programmen diesen Aspekt der flexiblen Veränderungsmöglichkeiten immer explizit zu machen.

Für einige *Spezialformen von Programmen* scheint die Übereinstimmung relativ groß zu sein:

- Kampagne (Nöcker 2011),
- Aktionsbündnis, Pakt, Netzwerk,
- Spezialprogramme für bestimmte Politik- oder Handlungsfelder,
- Settingbezogene Interventionsprogramme (allgemein von der Art des Settings unabhängig).

Am besten definiert ist die *Kampagne* als Gesamtplan aufeinander abgestimmter Maßnahmen mit Medienunterstützung (z. B. Fernsehspots, Plakataktionen), die meist längerzeitig, aber letztlich zeitlich begrenzt sowohl bezüglich der Dauer als auch der Handlungsbereiche angelegt sind.

Aktionsbündnis, Pakt, Netzwerk stehen als Überschriften über Programmen, die in systematischer Weise Kooperationen und die dafür nötigen Strukturen anregen sollen. Dabei sind sie meist auf spezifische inhaltliche Handlungsbereiche der Gesundheitsförderung festgelegt oder haben zeitlich begrenzte Fördermittel. Den Erfolg kann man an intermediären Parametern wie der Dauerhaftigkeit und der Nachhaltigkeit der geschaffenen Strukturen und der Erfüllung strategischer Zielsetzungen messen.

Spezialprogramme sind häufig primär dadurch bestimmt, in welchen *Politik- oder Handlungsfeldern* sie stattfinden. Das gilt zum Beispiel für die »medizinische Prävention« oder das Akti-

onsprogramm »Umwelt und Gesundheit«. In unserem eigenen Arbeitsfeld hatten wir es vor allem mit Spezialprogrammen für das Politikfeld »Gesundheit und Stadt« oder auch enger gefasst: für »Stadtentwicklung« zu tun. Settingbezogene Programme kann man allgemein und unabhängig von der Art des jeweiligen Settings folgendermaßen beschreiben: Es handelt sich um systematisch miteinander verknüpfte Maßnahmen zur Verbesserung von Lebens- und Arbeits*verhältnissen* sowie gesundheitsrelevanten *Verhaltens*weisen in einem abgegrenzten, sozial und räumlich definierten Umfeld, das heißt in einem bestimmten Ausschnitt der Lebenswelt von Menschen (einschließlich Organisation, Strukturen, Abläufen etc.). Einzelne settingbezogene Interventionen sind meist vereint in einem überregionalen Netzwerk von Settingprojekten desselben Settingtyps. Die bundesweiten Netzwerke sind wiederum meistens Teil von internationalen Netzwerken des jeweiligen Typs (also z. B. Health Promoting Hospitals, Healthy Cities etc.).

Sozialräumliche Settings wie »Stadt«, Quartier und ähnliches könnte man als »Settings 1. Ordnung« bezeichnen, da sie zahlreiche weitere Einzelsettings (Sub-Settings) enthalten (Schulen, Kindergärten, Betriebe und andere Settings 2. Ordnung).

Bei den *Programmen im Politikfeld »Gesundheit und Stadt«* muss man zwei Programmebenen unterscheiden: Erstens *Rahmen*programme der Gesundheitsförderung für mehrere Gebiete bzw. sozialräumliche Settings und zweitens Gesundheitsförderungsprogramme mit Sozialraumbezug für ein *(einzelnes) Gebiet*.

Rahmenprogramme, die nicht explizit primär für die Gesundheitsförderung angelegt sind, sondern z. B. aus der Stadtentwicklung oder der Umweltpolitik stammen, werden häufig als »integrierte Programme« bezeichnet, weil sie einen Sozialraum intersektoral, das heißt über mehrere Politiksektoren, breit gefächert, entwickeln wollen. Als Förderinitiative sind sie überregional. In der Regel versuchen sie auch in dem Sinne integriert zu handeln, dass Akteure aus kommunalen/staatlichen, zivilgesellschaftlichen und marktwirtschaftlichen Bereichen als Einzelakteure gewonnen werden sollen. Überregional heißt dabei nicht zwangsläufig bundesweit, wie es bei den Programmen Gesunde Stadt und Soziale Stadt der Fall ist, sondern oft handelt es sich um landesweite oder auf Metropolregionen bezogene Programme.

Ein weiteres besonders zu beachtendes, in der Regel (und so auch bei uns) vorhandenes Merkmal ist, dass die einzelnen Gebiete eines Rahmenprogramms danach ausgesucht sind, dass sie Stadtteile oder Quartiere *mit benachteiligten Bevölkerungsgruppen* sind, das heißt *Sozialräume mit besonderem Entwicklungsbedarf* (vgl. Milton et al. 2011).

Nach dieser Sondierung des gesamten Feldes von Interventionsprogrammen der Gesundheitsförderung und Prävention waren wir in der Lage, das spezielle Interventionsmuster in unserem Studiengebiet, dem Lenzviertel, präziser zu beschreiben. Dies ist allerdings nicht eindimensional möglich: Das spezielle Interventionsmuster »gesundheitsfördernde Quartiersentwicklung« vereint verschiedene Merkmale fast aller genannten komplexen Interventionstypen, besonders aber die Charakteristika

- eines integrierten sozialräumlichen Gesundheitsförderungsprogramms und
- eines integrierten Handlungsprogramms für Quartiersentwicklung.

Weitere wesentliche Merkmale sind die dynamische, »rollende« Planung und partizipative Weiterentwicklung mit den Bewohnern und Akteuren des Quartiers. Kusters et al. (2011, S. 101–104) machen in ihrer ohnehin sehr praxis- und nutzernahen Evaluationsanleitung dankenswerterweise in einer Gegenüberstellung von Patton (2011) sehr schön explizit, welches die Unterschiede sind zwischen der »traditional« und einer »complexity-sensitive developmental evaluation«.

Dass das Interventionsmuster für das Lenzviertel komplex ist, scheint uns offensichtlich zu sein. Weniger offensichtlich ist, wie eine darauf ausgerichtete komplexitätssensitive Gesamtevaluation aussehen kann. – Dazu müssen wir im nächsten Abschnitt erst konkreter beschreiben, aus welchen Bausteinen bzw. Teilinterventionen sich dieses Interventionsmuster zusammensetzt.

4 Programmbausteine der gesundheitsfördernden Quartiersentwicklung

Die Idee, die Bausteine der komplexen Intervention als »einfache Interventionen« zu bezeichnen, haben wir fallengelassen. Es stellte sich nämlich bald heraus, dass auch die kleinen, vermeintlich einfachen Interventionen oft die Kriterien der Komplexität erfüllen, die von maßgeblicher Seite in England zusammengestellt wurden (Craig et al. 2008):

- Menge und Schwierigkeit von Verhaltensweisen, die bei Akteuren/Leistungserbringern bzw. Empfängern/Patienten der Intervention vorausgesetzt/erwartet werden,
- Unterschiedliche Zielgruppen oder Organisationsebenen, auf die sich die Interventionen richten,
- Anzahl und Variabilität der Outcomes, sowie Grad der zugelassenen Flexibilität oder der eindeutigen Definition bzw. Manualisierung der Interventionen.

Tabelle 1
Seite 1 der Chronologie Lenzgesund (Auszug; Stand: 10.08.2010)

Jahr	Maßnahme	Beteiligte Einrichtungen *Status 2007*
1990–1994	Die Lenzsiedlung nimmt am EU-Programm Poverty III teil	
2000 Februar	Die Lenzsiedlung wird in das Hamburger Programm der Sozialen Stadtteilentwicklung aufgenommen	Lawaetz*, SAGA*, Bezirksamt, Lenzsiedlung e. V., StEB* *s. u. 2007/2*
	Das Gesundheitsamt Eimsbüttel beginnt mit der Planung von Angeboten für das Quartier	Gesundheitsamt, Lenzsiedlung e. V.
2000 September	Der allgemeine Runde Tisch macht ein Brainstorming zum Thema Gesundheit	15 Institutionen aus allen Bereichen; Gesundheit nur gering vertreten
2001 Mai	Eine Sprechstunde der Mütterberatung eröffnet unter dem Namen Babycafé in angemieteten Räumen im Bürgerhaus LS (1 x/Woche)	Gesundheitsamt, Lenzsiedlung e. V., einfal e. V.* *s. u. 2003/5*
2002 April	Start der »Gesundheitsgespräche für Frauen von Frauen« 7 Veranstaltungen in 2002 mit Übersetzungshilfen und Kinderbetreuung	Lenzsiedlung e. V., Gesundheitsamt, HAG e. V.*, ÄGGF e. V.*, Referentinnen aus diversen Einrichtungen des Sozial- u. Gesundheitsbereichs *s. u. 2003/3*
2002 September	Start einer qualitativen Befragung von Frauen zu ihrer gesundheitlichen Situation in der Lenzsiedlung und in ihrer Heimat (Werkauftrag)	Gesundheitsamt, Behörde für Umwelt und Gesundheit, Lenzsiedlung e. V. *abgeschlossen*
2002 Oktober	Die Familienhebamme eröffnet Büro und Sprechstunde im Bürgerhaus LS* und betreut vor allem Frauen aus der Lenzsiedlung (Werkauftrag)	Gesundheitsamt, BUG*, Lenzsiedlung e. V. *laufend*
2002 Dezember	Im Bürgerhaus LS* eröffnet ein Krabbeltreff (1 x/Woche) als Ergänzung zu den Angeboten der Mütterberatung und der Familienhebamme (Werkauftrag)	Gesundheitsamt, Lenzsiedlung e. V. *laufend*
2003 März	Zweiter Jahrgang »Gesundheitsgespräche für Frauen von Frauen« beginnt; 7 Veranstaltungen in 2003 mit externen Referentinnen, Übersetzungshilfen und Kinderbetreuung	Lenzsiedlung e. V., Gesundheitsamt, HAG e. V.*, ÄGGF e. V.*, Rauhes Haus*, Referentinnen aus diversen Einrichtungen des Sozial- u. Gesundheitsbereichs *ab 2004/1 kursorische Fortsetzung der Reihe*
2003 April	Psychomotorisches Turnen (1–2 x/Woche) für Kinder aus der Kita Vizelinstraße und aus einer offenen Gruppe des Kinderclubs Lenzsiedlung e. V.	Kita Vizelinstraße, Lenzsiedlung e. V. *zzt. nicht*
2003 Mai	Die Mütterberatung schließt ihre Sprechstunde in der Lenzsiedlung wegen geringer Nutzung und Sparauflagen aus dem politischen Raum	
2003 Mai	Impfaktion in der LS* im Rahmen der Nationalen Impfwoche	Gesundheitsamt, Lenzsiedlung e. V.

* Auf die Erläuterungen der Abkürzungen wurde hier verzichtet.

Wir haben diese Bausteine daher etwas umständlich als »weniger einfache Interventionen« benannt. Mit unserer Zurückhaltung bezüglich des Begriffs »einfache Interventionen« liegen wir auf einer Linie mit den Autoren des Medical Research Council (Craig et al. 2008), die in ihrer Definition von komplexen Interventionen auch ausdrücklich erwähnen:

- no sharp boundary between simple and complex interventions,
- few interventions are truly simple, but the number of components and range of effects may vary widely.

Grundlage unserer Identifizierung der Bausteine und einer Klassifikation dieser Bausteine war die Chronologie der Aktionen und Interventionen im Lenzviertel. Eine frühe Fassung ist nachzulesen in der Quartiersdiagnose unseres Projekts (Kohler et al. 2007). Um dies anschaulich zu machen, geben wir eine Seite der Chronologie als Facsimile wieder (Tabelle 1).

Die Entwicklung unserer Klassifizierung von gemeinwesenorientierten Interventionen nach Typen und Untertypen geschah nach der Methode der Grounded Theory: Ausgehend von den in der Chronologie enthaltenen Aktivitäten wurden Gruppen mit möglichst homogenen Merkmalen gebildet, mit einer Überschrift versehen und anhand der Merkmale charakterisiert. Das Ergebnis wurde mehrfach von unserer Projektgruppe diskutiert und verbessert[1]. Trotz dieser gründlichen »Arbeit am Begriff« sind die Typen und Untertypen als erste Versuche einer systematischen Unterteilung zu verstehen, als "Work in Progress". Ergänzungen, Veränderungen und genauere Definitionen sollten unseres Erachtens in einem größeren Rahmen als dem eines einzelnen Projektes erfolgen. Zunächst stellen wir fünf Typen vor, um daran anschließend nochmals für jeden Typ die Untertypen zu benennen.

Fünf Typen im *Überblick*:

- Aktionen (mit 4 Untertypen): an Zielgruppen/Endadressaten gerichtet, einmalig oder kurzfristig; auf Bewusstseinsbildung, Motivierung, Aktivierung gerichtet; können an einem Aktionsort oder »multilokal« stattfinden,
- Angebote (mit 4 Untertypen): ebenfalls an Zielgruppen/Endadressaten gerichtet, mehrmalig oder langfristig, auf direkte Hilfe oder Kompetenzerweiterung bei den Zielgruppen ausgerichtet,
- Programmentwicklung (mit 4 Teilbereichen/Untertypen): essentiell wichtig bei dynamischen und partizipativen Interventionsprogrammen; als Interventionstyp nicht nötig bei vorab festgelegten weniger komplexen Interventionen,
- Strukturentwicklung (mit 4 Teilbereichen/Untertypen): im Idealfall eng verknüpft mit Programmentwicklung; jedes Programmelement sollte durch Kompetenz- und Strukturentwicklung (Capacity Buildung; vgl. Trojan, Nickel 2011) nachhaltig abgesichert werden,
- Evaluation (mit 3 Teilbereichen/Untertypen): lässt sich auf der Handlungsebene analytisch zwar trennen, ist aber im Idealfall bei dynamischen und partizipativen Programmen voll integriert.

Als *Untertypen* bei den *Aktionen* haben wir identifiziert:

- offene Infoveranstaltung, das heißt jeder kann hingehen, Erfolg misst sich stark an der Zahl der Teilnehmer/-innen,
- Aktionstag/-woche, das heißt keine einmalige, sondern mehrfache Aktionen mit deutlich größerer Sichtbarkeit, aber auch schwierigerer Evaluierbarkeit,
- (lokale) Kampagne: wie eine große Kampagne, jedoch kleinräumiger und von der Wahl der Medien her einfacher angelegt,
- Freizeitveranstaltung, sehr geeignet um im »Huckepackverfahren« (auch »Andocken« genannt) Anliegen der Gesundheitsförderung größeren Gruppen der Bevölkerung nahe zu bringen.

Im Bereich der *Angebote* wurden die folgenden vier *Untertypen* identifiziert:

- Sprechstunde/offenes Beratungsangebot, das heißt in der Regel nur einmalige Nutzung und entsprechend vermutlich geringere Effekte,
- Offenes Gruppenangebot, das heißt, viele können öfter kommen, müssen es aber nicht,
- Aufsuchende/niedrigschwellige Hilfe, das heißt im Gegensatz zu den beiden vorangegangenen

1 Ich möchte an dieser Stelle ganz herzlich unserem Praxispartner Christian Lorentz sowie den Projektmitgliedern Waldemar Süß, Stefan Nickel, Karin Wolf für die kreativen Diskussionen danken.

»Komm«-Angeboten eine zugehende Hilfe, z. B. die Familienhebamme,
- Kurs, das heißt strukturiertes Angebot mit mehreren Terminen, bei denen eine bestimmte Zahl von Teilnehmerinnen bis zum Ende dabei ist.

Programmentwicklung hatte ebenfalls 4 unterscheidbare *Untertypen*:
- thematische Prioritätensetzung, was auf verschiedenen Ebenen oder bei verschiedenen Akteuren erfolgen könnte, im Lenzviertel aber in der Regel als ergebnisorientierte Diskussion am Runden Tisch ablief,
- Themen-Integration/Andocken; dabei ging es häufig um das Prinzip, den Zugang zu bestimmten Angeboten oder Aktionen dadurch zu erleichtern, dass Gesundheitsförderung z. B. im Rahmen von Geselligkeits- und Freizeitveranstaltungen stattfand,
- Programmplanung, in der Regel als Aktivität des Runden Tisches,
- Programmfestlegung/Revision, ebenfalls in der Regel am Runden Tisch.

Strukturentwicklung ließ 5 *Untertypen* erkennen:
- Vernetzungsmaßnahme, d. h. die Initiierung von Kooperation mit anderen Akteuren,
- Koordinierungsmaßnahme, d. h. die Regelung vorhandener Kooperation,
- AG/AK-Gründung, d. h. die Neugründung eines Zusammenschlusses für die kooperative themen- oder aufgabenbezogene Bearbeitung eines Handlungsfeldes,
- Infrastrukturbildung, d. h. die Institutionalisierung einer neuen Quelle für zusätzliche personelle und/oder ökonomische Ressourcen,
- (Akteurs-)Fortbildung, d. h. die Verankerung zusätzlicher Kompetenzen bei relevanten Akteuren.

Während die ersten 3 dieser Untertypen auf die Entwicklung von *Programm*-Strukturen ausgerichtet sind, also notwendige strukturelle Ergänzungen des Typs Programmentwicklung darstellen, sind die beiden letzten Untertypen als *unmittelbare* Kapazitätsentwicklung, man könnte sagen, als Capacity Buildung im engeren und eigentlichen Sinn anzusehen.

Evaluation zeigte drei *Untertypen*:
- Situationsanalyse/Bestandsaufnahme, in unserem Fall relativ umfangreich und aufwendig, aber im Prinzip als Teil des Public Health Action Cycle als Beginn bei *allen* Interventionen nötig,
- Evaluationsmaßnahme, z. B. eine Befragung der Bewohner oder von Akteuren; wird als Intervention gezählt, weil sie immer auch aufklärerischen und häufig aktivierenden Charakter hat,
- Evaluationsrückmeldung, wobei es sich meistens um die externe Rückmeldung des Forschungsprojektes handelte; könnte in Projekten mit weniger Evaluationsressourcen auch einfach der Bericht am Runden Tisch oder in ähnlichem Gremium über den wahrgenommenen Erfolg oder Misserfolg eines Angebots oder einer Aktion sein.

Tatsächlich haben sich die 5 Typen mit ihren Untertypen fast »organisch« aus der systematisierenden Zusammenfassung der so genannten Chronologie der Maßnahmen im Lenzviertel ergeben. Beim genaueren Hinsehen lässt sich diese rein induktiv gewonnene Klassifikation jedoch auch theoretisch und systematisch genauer verorten. Wir können drei Ebenen von Interventionen unterscheiden:
- Ebene 1: *Aktionen und Angebote* sind vorrangig *an die Endadressaten* gerichtet;
- *Aktionen* sind im Rahmen eines Lebensweisenkonzepts (Abel, Ruckstuhl 2011) vorrangig zu verstehen als auf gesundheitsbezogene *Orientierungen und Verhaltensweisen* gerichtet; *Angebote* sind gerichtet auf die individuellen Ressourcen, das heißt die *personengebundenen Faktoren des Ressourcenkonzepts*.
- Ebene 2: betrifft Programmentwicklung und Strukturentwicklung. Beide sind als *Steuerungs- bzw. (Infra-)Struktur-Interventionen* zu verstehen, also einerseits auf die Steuerung dessen, was auf Ebene 1 passiert und andererseits auch auf die *Verbesserung der sozialen Ressourcen*, d. h. die Lebensverhältnisse mit ihren Infrastrukturen und die "Capacities" gemäß dem Lebensweisenkonzept.
- Ebene 3: ist die Ebene der Interventionen für *Qualitätsentwicklung* auf den Ebenen 1 und 2.

Ebene 1 kann man auch als *direkte* Basisinterventionen für und mit den Zielgruppen bzw. Endadressaten bezeichnen. Ebene 2 sind die indirekten, das heißt den Zielgruppen nur *vermittelt* zugutekommenden Interventionen. Die 3. Ebene der Evaluation kommt *nochmals indirekter*, nämlich erst durch Qualitätsentwicklung auf den Ebenen 1 und 2, den eigentlichen Zielgruppen zugute.

Eine weitere (vierte) Interventionsebene, die aus methodischen Gründen nicht aufscheint (weil in der Chronologie der Interventionen *im* Lenzviertel nicht angesprochen), wären Interventionen in den Kontext außerhalb des Lenzviertels. Zu diesem Kontext gehören beispielsweise die Entwicklung des Hamburger Programms RISE (Rahmenprogramm integrierte Stadtentwicklung) oder des Paktes für Gesundheit der Gesundheitsbehörde, die beide auch für die Erhaltung und Weiterentwicklung der Aktivitäten im Lenzviertel förderlich sein können. Dazu gehören aber auch parallele Programme der Stadt Hamburg, die die Ressourcen für die bezirklichen Ebenen stark einschränken und sich auf diese Weise negativ auf die bisherigen Maßnahmen des Gesundheitsamtes Eimsbüttel auswirken werden.

Zu fordern ist darüber hinaus noch eine 5. Ebene: die Evaluation der Evaluation und Qualitätsentwicklung unter Berücksichtigung aller vier vorher genannten Ebenen. Nur wenn wir schlüssig sagen können, dass die Evaluation und die daraus resultierende Qualität einen hohen Standard haben, lassen sich die Ergebnisse der Evaluation nutzen für eine Verbesserung der allgemeinen Evidenzlage bezüglich des Ansatzes quartiersbezogener Gesundheitsförderung. Tatsächlich gibt es ja auch Qualitätsstandards für wissenschaftliche Evaluationen im Allgemeinen und für die Evaluation von gesundheitsfördernden Interventionen im Besonderen. Allerdings kann man bisher nicht sagen, dass die vielen positiven bis euphorischen Berichte über Projekte und Programme in systematischer Weise darauf hingeprüft wurden, wie groß ihre Aussagekraft für eine systematische Bilanzierung der Evidenzlage ist.

Abschließend möchte ich versuchen, die für komplexe Setting-Interventionen idealerweise notwendigen Interventionselemente bzw. Interventionsebenen in einer schematisierenden Tabelle zusammenfassen. Die Kategorisierung und Begrifflichkeit der Interventions*ziele* orientiert sich dabei am Lebensstil-/Lebensweisenkonzept (Abel, Ruckstuhl 2011), das im Kern folgendermaßen charakterisiert wird:

Gesundheitsrelevante Lebensstile sind typische Muster von
- ◇ gesundheitsrelevanten Verhaltensweisen (»für die Gesundheit direkt relevant«)
- ◇ gesundheitsbezogenen Orientierungen (»Werte, normative Vorgaben und Einstellungen«)
- ◇ gesundheitsrelevanten Ressourcen, und zwar

a) individuellen Ressourcen: personengebundenen Faktoren: Wissen und Fähigkeiten, persönliche Mittel;
b) sozialen Ressourcen: außerhalb des Individuums, Lebensverhältnisse: Familie, Freunde; Wohnbedingungen, Existenz sichernde Arbeitsplätze, bewegungsförderliche Infrastruktur etc.

Die Tabelle 2 stellt einen ersten Versuch dar, "work in progress" also, mit der Absicht, so weit wie möglich die ansonsten nur diffus beschreibbare Setting-Intervention in ihre Bestandteile zu zerlegen. Wie weit die einzelnen Elemente stimmig sind und ob sie mehr sein können als eine Checkliste muss die weitere Diskussion zeigen.

Die Tabelle 2 repräsentiert in gewisser Weise eine Hierarchie von Interventionen und daraus folgend auch von Evaluationsebenen, die *idealtypisch* zu einer vollständigen komplexen Gesundheitsförderung in einem Setting dazugehören:
- ▷ die Basisinterventionen (Aktionen und Angebote), unmittelbar an die Endadressaten gerichtet;
- ▷ auf die nachhaltige Verbesserung von Kompetenzen und Strukturen im Setting gerichtete Kapazitätsentwicklung,
- ▷ Interventionen, die das Programm selbst bzw. seine Steuerungsakteure adressieren,
- ▷ Kontext-Interventionen und schließlich die
- ▷ praxis- und wissenschaftsangemessene Verbesserung der Qualitätsentwicklung und Evaluation selbst, als letztendlichen Garanten für die kontinuierliche Verbesserung der Evidenzlage für Interventionen auf den vorher genannten Ebenen.

Tabelle 2
Kategorisierung notwendiger Interventionstypen in einer komplexen Setting-Intervention

Interventionstyp	Adressaten/Zielgruppen	Interventionsziel(e)
Aktionen	Bevölkerung allgemein; »Endadressaten«	Orientierungen/Normen für Verhaltens- und Verhältnisziele verändern
Angebote	Bevölkerung; spezifische Zielgruppen; »Endadressaten«	Individuelle Ressourcen für Verhaltens- und Verhältnisänderungen verbessern; individuelles Empowerment
Kapazitätsentwicklung (im engeren Sinne)	Akteure, (operative Durchführungsebene)	Soziale Ressourcen/Verhältnisse im Setting verändern; Community Empowerment
Programmentwicklung	Akteure (Steuerungsebene)	Partizipative Entwicklung und Verbesserung von Strukturen, Prozessen, Umsetzung und Evaluation des Programms steuern; Leadership Empowerment
Kontextbeeinflussung	Akteure außerhalb des Settings und des Programms	Voraussetzungen und Rahmenbedingungen im Politik- und Praxisraum außerhalb des Settings beeinflussen
Qualitätsentwicklung und Evaluation der Evaluation	Qualitätsverantwortliche und Evaluatoren innerhalb und außerhalb des Programms	Verbesserung von Qualitätsentwicklungs- und Evaluationspraxis; Wissenschaftsentwicklung; Evidenzbildung

Eigentlich ist die Vorstellung verschiedener, hierarchisch gestufter Ebenen aber etwas irreführend. Viel besser passt das Bild der Zahnräder eines Uhrwerks, die alle zusammengenommen die Intervention antreiben, auch wenn sie unterschiedlich weit entfernt sind von den Zeigern (dem Verhalten und den Verhältnissen), die letztendlich bewegt werden sollen.

5 Mehr Evidenz aufbauen für Interventionen der Gesundheitsförderung

Wir sind in unserem Forschungsprojekt der Meinung, dass eine Strukturierung und einheitliche Benennung der Bausteine eines komplexen Gesundheitsförderungsprogramms eine unbedingt *notwendige* Voraussetzung ist für qualitativ hochstehende Evaluationen und Aussagen zur Wirkung von Gesundheitsförderungsinterventionen.

Eine *hinreichende* Voraussetzung ist die Festlegung einer Terminologie und Klassifikation allerdings nicht. Gern würden wir dazu schon weiter gehende Angaben machen können. Aber in diesem Punkt müssen wir uns leider eng auf das gestellte Thema beschränken: Wir können zu diesem Zeitpunkt kein fertiges Rezept liefern, wie die schlüssige Gesamtevaluation des Präventionsprogramms Lenzgesund so gestaltet werden kann, dass sie verallgemeinernde Aussagen über das Interventionsmuster »gesundheitsfördernde Quartiersentwicklung« zulässt. Allerdings ist bei der Diskussion über diese Fragen eine Vision entstanden, was wir für »evidenzbasierte« bzw. wissenschaftlich abgesichert vielversprechende (promising) Gesundheitsförderungsinterventionen bräuchten.

Plakativ zusammengefasst müssten wir, die Community der Gesundheitsförderer, die vorhandene web-basierte Projektdatei des Kooperationsverbundes Gesundheitliche Chancengleichheit ergänzen

- um eine *nach Typen geordnete, strukturierte Interventionsdatei,*
- mit gesammeltem Wissen über *Kontexte von Gelingen und Scheitern,*
- in Form halbstandardisierter *Akteurs- bzw. Anwenderbewertungen,*
- resultierend in regelmäßig aktualisierten *Interventionsprofilen,*
- mit Beschreibung des *generalisierbaren Kerns* der komplexen Intervention,
- plus flexiblen, *kontextabhängigen variablen Anteilen,*
- mündend in eine differenzierende Bilanz, was *unter Alltagsbedingungen* wirkt, also nicht nur unter Idealbedingungen (wie z. B. in Modellprojekten und ähnlichem).

Die Vision wäre also eine Datei mit dem gesammelten strukturierten Wissen über die Evidenzlage

bei komplexen und weniger komplexen Interventionen. Der erste Schritt der Evidenzproduktion wäre im Idealfall eine Bilanz der Ausgangslage nach Art von Reviews. Jeder Nutzer müsste danach dann verpflichtet sein, seine Erfahrungen mit den jeweiligen Interventionen wiederum in diese Datei einzubringen und damit zu einer erneuerten Synthese des vorhandenen Wissens beizutragen. Der Kreislauf von Evidenzproduktion, -nutzung und erneuter -synthese entspricht im Prinzip dem Vorgehen wie im Public Health Action Cycle und gleicht auch der von Töppich vorgeschlagenen Qualitätsentwicklungsspirale (s. BVPG 2011, S. 6) bzw. dem von Wright et al. (2012) vorgeschlagenen »Evidenz-Zyklus«.

Es ist zu wünschen, dass die Beiträge dieses Bandes uns auf dem Weg zu diesen weitreichenden Zielen ein Stück weiter bringen!

Literatur

Abel T, Ruckstuhl B (2011) Lebensweisen/Lebensstile. In: BZgA (Hrsg) Leitbegriffe der Gesundheitsförderung und Prävention. Glossar zu Konzepten, Strategien und Methoden. 4. Auflage. Verlag für Gesundheitsförderung, Gamburg, S 365–368

BVPG/Bundesvereinigung Prävention und Gesundheitsförderung (2011) Statusbericht 4: Dokumentation der Statuskonferenz 2011 zu »Qualitätsentwicklung in Prävention und Gesundheitsförderung«. Eigenverlag www.bvpraevention.de (Stand: 07.03.2012)

Campbell M et al. (2000) Framework for design and evaluation of complex interventions to improve health. BMJ 2000, 321: 694–696

Craig P et al. (2008) Developing and evaluating complex interventions: new guidance. Prepared on behalf of the Medical Research Council www.mrc.ac.uk/complexinterventionsguidance (Stand: 07.03.2012)

Gesundheit BB e. V. und BZgA http://www.gesundheitliche-chancengleichheit.de -> Glossar (Stand: 12.03.2012)

Härter M (2011) Komplexe Interventionen, Definition und Evaluationskonzepte. Referat bei der DNEbM Akademie – LEUCOREA. Wittenberg, 9. September 2011

Kohler S, Mossakowski K, Süß W et al. (Hrsg) (2007) Beiträge zur Quartiersdiagnose. Kindergesundheit in der Lenzsiedlung. Eigenverlag, Hamburg

Kusters C, van Vugt S, Wigboldus S et al. (2011) Making Evaluations Matter: A Practical Guide for Evaluators. Centre for Development Innovation, Wageningen University & Research Centre, Wageningen, The Netherlands www.cdi.wur.nl (Stand: 07.03.2012)

Milton B, Moonan M, Taylor-Robinson D et al. (Eds) (2011) How can the health equity impact of universal policies be evaluated? Insights into approaches and next steps. Synthesis of discussions from an Expert Group Meeting. A joint publication between the WHO European Office for Investment for Health and Development, Venice, Italy and the Liverpool WHO Collaborating Centre for Policy Research on Social Determinants of Health. Held at the University of Liverpool, 2–4 November 2010

Mossakowski K, Nickel S, Süß W et al. (2010) Die Quartiersdiagnose: Daten und Ansätze für ein stadtteilorientiertes Präventionsprogramm des Öffentlichen Gesundheitsdienstes – ausgewählte Ergebnisse eines Forschungsprojektes. In: Laverack G (Hrsg) Gesundheitsförderung & Empowerment. Verlag für Gesundheitsförderung, Gamburg, S 115–127

Nöcker G (2011) Gesundheitskommunikation und Kampagnen. In: BZgA (Hrsg) Leitbegriffe der Gesundheitsförderung und Prävention. Glossar zu Konzepten, Strategien und Methoden. 4. Auflage. Verlag für Gesundheitsförderung, Gamburg, S 291–297

Patton M (2011) Developmental Evaluation: Applying Complexity Concepts to Enhance Innovation and Use. Guilford Press, New York

Töppich J, Linden S (2011) Evaluation. In: BZgA (Hrsg) Leitbegriffe der Gesundheitsförderung und Prävention. Glossar zu Konzepten, Strategien und Methoden. 4. Auflage. Verlag für Gesundheitsförderung, Gamburg, S 69–72

Trojan A, Nickel S (2011) Capacity Building/Kapazitätsentwicklung. In: BZgA (Hrsg) Leitbegriffe der Gesundheitsförderung und Prävention. Glossar zu Konzepten, Strategien und Methoden. 4. Auflage. Verlag für Gesundheitsförderung, Gamburg, S 42–44

Trojan A, Süß W (2011) Gesundheitsbezogene Gemeinwesenarbeit. In: BZgA (Hrsg) Leitbegriffe der Gesundheitsförderung und Prävention. Glossar zu Konzepten, Strategien und Methoden. 4. Auflage. Verlag für Gesundheitsförderung, Gamburg, S 122–124

Wright M, Kilian H, Brandes S (2012) Praxisbasierte Evidenz in der Prävention und Gesundheitsförderung bei sozial Benachteiligten. Erscheint in: Das Gesundheitswesen

Wirkungen und Wirkungsnachweis bei komplexen Interventionen

Wolfgang Bödeker

1 Einleitung

Das Thema komplexe Interventionen erfreut sich derzeit in Deutschland regen wissenschaftlichen Interesses. Der Bericht von der 13. Jahrestagung des Deutschen Netzwerks Evidenzbasierte Medizin (DNebM) e. V. 2012 etwa fasst die Aufgabe wie folgt zusammen: »Der Nutzen eines Arzneimittels lässt sich in randomisiert-kontrollierten Studien nachweisen. Schwieriger wird diese Evidenzgenerierung, wenn es um Maßnahmen geht, die sich aus mehreren Einzelkomponenten zusammensetzen, … Viele Einzelkomponenten müssen beobachtet und bewertet werden, deren Interaktionen und Einflüsse auf den Gesamtnutzen der untersuchten komplexen Interventionen häufig nicht klar sind. Seit Jahren wird daher zur Bewertung und Synthese von komplexen Interventionen eine Differenzierung der methodischen Verfahren gefordert.«

Der Workshop, der dieser Publikation zu Grunde liegt, folgte einer etwas anderen Programmatik. In der seinerzeitigen Einladung hieß es »In Deutschland gibt es eine Reihe von komplexen Interventionsprogrammen in der Prävention, in der unter einem gemeinsamen organisatorischen Dach viele Akteure mit verschiedenen Methoden an verschiedenen inhaltlichen Präventionsaspekten arbeiten. Während die Evaluation von Einzelmaßnahmen auf bewährte Instrumentarien zurückgreifen kann, gibt es zu Evaluation komplexer Interventionsprogramme bisher kaum methodische Reflexionen.«

Die Zugänge gleichen sich in der Einschätzung, dass Effekte von Einzelmaßnahmen methodisch angemessen beurteilt werden können. Während aber das DNebM das Problem komplexer Interventionen bei der Synthese der Einzelergebnisse, also der Evidenzbasierung, sieht, liegt dem Workshop die Vermutung zugrunde, dass schon die Evaluation problematisch ist. Evaluation und Evidenzbasierung sind aber nicht dasselbe. Evaluation hat ihre Rolle in der Beurteilung, ob gesetzte Ziele durch ein Projekt oder Programm erreicht wurden. Sie umfasst grundsätzlich mehr als Wirkungsevaluation, da Ziele auch mit Blick auf die Akteure und die Instrumente einer Intervention gefasst sein können und daher ggf. durch eine einfache Beschreibung des Erreichten evaluiert werden können. Wirkungsevaluation dagegen zielt auf die Beurteilung, ob ein gewünschter Effekt durch die gewählte Intervention ausgelöst wurde. Wirkungsevaluation unterliegt damit – unabhängig von der Art der Intervention – höheren methodischen Anforderungen, weil es nicht nur um die Deskription von Ereignissen, sondern um deren kausale Attribuierung geht. Wirkungsevaluation ist damit methodisch gleichwertig beispielsweise zu einer klinischen Therapiestudie, obwohl hier kaum der Begriff Evaluation benutzt würde. Evidenzbasierung schließlich setzt die Wirkungsevaluation mehrerer gleichartiger Interventionsstudien voraus, denn aus deren Zusammenfassung und Beurteilung wird die Evidenzbasis gebildet.

In dem folgenden Beitrag wird zunächst betrachtet, wie Wirkungsaussagen grundsätzlich erzeugt werden und in welchem Rahmen sie als aussagekräftig betrachtet werden. Die Reflexion des Verständnisses von komplexen Interventionen führt dann zur Prüfung, ob diese besondere Herausforderungen an den Wirkungsnachweis, mithin also an die Wirkungsevaluation stellen.

2 Wirkungsmodelle und Wirkungsnachweis

Mit einer Wirkungsevaluation wird der Frage nachgegangen, ob Ereignisse Folge einer Intervention sind. Wirkungsevaluation zielt auf das Aufdecken der Ursachen des Ereignisses und betrifft damit eine Kernfrage der Erkenntnistheorie.

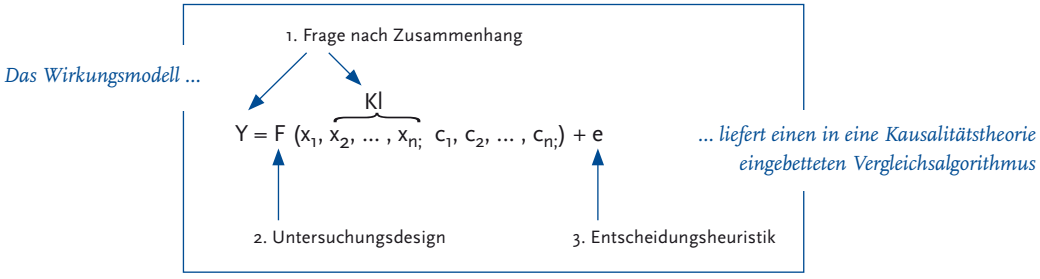
Als Grundlage der Wirkungsevaluation wird im Allgemeinen von einem schlichten Wirkungsmodell mit drei konstitutiven Elementen ausgegangen (Abb. 1). Erstens steht im Vordergrund die das Wirkungsmodell evozierende Untersuchungsfrage. Zweitens wird diese in ein Untersuchungsdesign (F) umgesetzt, das eine Quantifizierung des Zusammenhangs zwischen der Intervention (KI) und dem Ereignis/Effekt (Y) ermöglicht. Ein

Unterschied zwischen dem Effekt nach Intervention und ohne Intervention soll dann die ursächliche Rolle der Intervention beweisen. Der Nachweis wird dadurch verkompliziert, dass neben der Intervention weitere »Einflussfaktoren« als Konstituenten der Ursache-Wirkungsbeziehung nicht ausgeschlossen werden können (x_i, c_i). Diese Einflussfaktoren können ihrerseits Ursachen oder aber konfundierende oder moderierende Faktoren darstellen, müssen also von der interessierenden »wahren« Ursache differenziert werden, damit die beobachteten Unterschiede ursächlich der Intervention attribuiert werden können. Durch das Untersuchungsdesign sollen die Effekte dieser zusätzlich zur Intervention bestehenden Einflüsse erkennbar oder durch eine gleichmäßige Verteilung in der Interventions- und Kontrollgruppe ausgeschlossen werden (z. B. durch Randomisierung). Drittes konstitutives Element des Wirkungsmodells ist schließlich eine Entscheidungsheuristik. Die Beobachtung eines Effektunterschieds führt nämlich nicht automatisch zur Entscheidung über das Vorliegen eines solchen. Es bedarf daher einer Regel, wonach z. B. 15 % Effekt bei einer Interventionsgruppe und 10 % bei der Kontrollgruppe ein Unterschied sein soll, während davon etwa bei 12 % gegenüber 10 % nicht ausgegangen werden kann. Eine Entscheidungsheuristik dient also der Quantelung der Quantifizierung, wofür oft ein statistischer Test herangezogen wird.

Das grundlegende Wirkungsmodell wird in einer Vielzahl von methodischen Variationen verfeinert, die aus der Diskussion technischer Fragen folgen, etwa: welches Untersuchungsdesign ist angemessen, welche Einflussfaktoren sind wie zu erheben, welche funktionalen Zusammenhänge zwischen den Einflussfaktoren sollen angenommen werden, welcher statistische Test ist angemessen. Diese Variationen verändern aber nicht den Grundtypus des Wirkungsmodells und seiner impliziten erkenntnistheoretischen Prinzipien, die den Wirkungsnachweis auf Basis dieses Modelles gegebenenfalls ermöglichen. Einerseits bedarf es nämlich des sogenannten Kausalitätsprinzips, wonach Ereignisse nicht ohne dazugehörende Ursache auftreten und andererseits dem Determinismusprinzip, wonach Ursachen ihre Wirkungen eindeutig hervorbringen, gleiche Ursachen also gleiche Wirkungen determinieren. Diese beiden Grundpfeiler einer Wirkungsevaluation sind indes erkenntnistheoretisch umstritten und alle gegenwärtigen wissenschaftlichen Kausalitätstheorien stimmen hierin nicht überein (vgl. Baumgärtner 2007). Geht man zudem von dem in der Einleitung noch vage gehaltenen Begriff der komplexen Intervention aus, so ergibt sich eine weitere Erschwernis des Erkenntnisprozesses. Nämlich auch die Verzahnung der Komponenten einer Intervention und die Auffassung, Ursachen würden sich in einer Wirkungskette fortpflanzen (Transitivität) steht auf keiner unzweifelhaften erkenntnistheoretischen Grundlage. Denn letztlich käme damit nur noch die Geburt einer Person als Ursache für ihren Tod in Frage.

Übereinstimmung besteht bei den gängigen Kausalitätstheorien indes darin, dass eine Instanz sich nicht selbst verursacht, Selbstverursachung also kein sinnvoller Begriff für eine Ursache-Wirkungsbeziehung ist. Damit ist auch klar, warum das grundlegende Wirkungsmodell der Abbildung 1 recht schlicht beschrieben werden kann. »Kausale Netze« lassen sich zwar auch einfach konstruieren, führen aber zur Selbstverursachung der vernetzten Komponenten, weshalb hier mit den üblichen

Abbildung 1
Das grundlegende Wirkungsmodell einer Evaluation



Kausalitätstheorien die Frage nach Ursache und Wirkung nicht zu beantworten ist.

Zusammenfassend kann also hervorgehoben werden:

1. Das dargestellte grundlegende Wirkungsmodell eignet sich als Blaupause für Nachweisverfahren in der Evaluation. Es kann methodisch vielfältig variiert werden, um zu einer möglichst verzerrungsarmen Interventionsstudie zu gelangen. Dabei ändert sich jedoch nicht der epistemische Geltungsanspruch. Das Ziel, »verursachende« Faktoren auf Basis eines Wirkungsmodells zu identifizieren, impliziert, dass zur Beschreibung des »intervenierten Systems« nicht alle Konstituenten erforderlich sind. Wirkungsevaluation ist daher grundsätzlich ein reduktionistisches Vorhaben und basiert einerseits auf der Annahme, Faktoren/Prozesse separieren zu können und andererseits, diese von der Wahrnehmung der Evaluatoren unabhängig halten zu können.
2. Das Streben nach verzerrungsarmen Studiendesign hat seine epistemische Basis in einem Kausalitätsmodell. Den meisten Diskutanten dürfte vertraut sein, dass Kausalitätsüberlegungen »schwierig« sind. Dass es aber keine allgemein akzeptierte Theorie der Kausalität gibt, sondern auch hier die aus den empirischen Wissenschaften bekannten Schulenstreits ausgetragen werden, wird für viele überraschend sein. Und auch, dass man sich mit der Wertschätzung eines bestimmten Verständnisses von komplexen Intervention ggf. schon im Widerspruch zu einer ansonsten geschätzten Kausalitätstheorie begeben hat.

3 Komplexe Interventionen und komplexe Systeme

3.1 Was wird unter komplexen Interventionen verstanden?

In Deutschland wird gegenwärtig in verschiedenen Wissenschaftsdisziplinen über komplexe Interventionen diskutiert. Auffällig ist, dass die Erörterungen weitgehend ohne eine genaue Definition des Gegenstands auszukommen scheinen. Oftmals wird aber auf eine Veröffentlichung des englischen Medical Research Council (MRC), bereits aus dem Jahre 2000, Bezug genommen. Hierin werden komplexe Interventionen verstanden als “a number of separate elements which seem essential to the proper functioning of the interventions although the ‘active ingredient’ of the intervention that is effective is difficult to specify” (MRC 2000). Die Veröffentlichung liegt inzwischen in einer neuen Auflage vor, in der die Definition verbreitert wird: “Complex interventions are usually described as interventions that contain several interacting components. There are however, several dimensions of complexity: it may be to do with the range of possible outcomes, or their variability in the target population, rather than with the number of elements in the intervention package itself. It follows that there is no sharp boundary between simple and complex interventions. Few interventions are truly simple, but there is a wide range of complexity.... Some dimensions of complexity

- Number of and interactions between components within the experimental and control interventions
- Number and difficulty of behaviors required by those delivering or receiving the intervention
- Number and variability of outcomes
- Degree of flexibility or tailoring of the intervention permitted.” (MRC 2008)

Die Revision führt gegenüber der früheren Begrifflichkeit zu einer Konkretion insofern, als dass es nun nicht mehr lediglich um eine Vielzahl, sondern um interagierende Elemente geht. Gleichzeitig wird das Verständnis von Komplexität auf die betrachteten Effekte und sogar auf die Subjekte der Intervention erweitert. Folge der unscharfen und vieldimensionalen Definition ist, dass eine

Unterscheidung zwischen einfachen und komplexen Intervention kaum möglich ist.

Das Begriffsverständnis des MRC findet sich auch in der gegenwärtigen Debatte in Deutschland wieder. Mühlhauser et al. (2011) etwa greifen das MRC-Verständnis auf: »Viele medizinische Maßnahmen sind komplexe Interventionen. Sie bestehen aus mehreren Einzelkomponenten, die sich wechselseitig bedingen und ihrerseits in komplexe Kontexte implementiert werden. Beispiele dafür sind Stroke Units, Disease-Management-Programme oder Projekte zur Verbesserung der Krankenhaushygiene. Ähnlich komplexe Interventionen gibt es in assoziierten Berufs- und Handlungsfeldern. Zum Beispiele Prävention von Sturz und Dekubitus in der Pflege, Ernährungs- und Sportprogramme in Schulen …«. Sie beziehen damit in ihre Definition Interventionen ein, die wie etwa die Sturz- und Dekubitusprävention in klassischen Studiendesign mit klassischen Studientypen wie RCT untersucht und in systematischen Reviews bewertet worden sind (Cameron et al. 2010, McInnes et al. 2011). Nach dem in der Einleitung wiedergegebenen Verständnis des DNebM entsprechen sie mithin dem Typus der einfachen Intervention. Auch hier erfolgt also keine klare Grenzziehung zwischen einfachen und komplexen Interventionen.

In der Gesundheitsförderung wird von einem anderen Verständnis ausgegangen, wonach offenbar einfache Interventionen per se nicht dem Kanon idealtypischer Maßnahmen entsprechen. In den BZGA-Leitbegriffen wird an der Einschätzung festgehalten, dass »Ein Evidenzbegriff, der wie in der Medizin so eng verknüpft ist mit dem naturwissenschaftlichen Experiment, ist für die Gesundheitsförderung fragwürdig. Die RCT gilt dort als unangemessen, ja sogar kontraproduktiv. Dementsprechend wurde vorgeschlagen, Evidenz in der Gesundheitsförderung als umfassenderes, plausibles Wissen über die Wirksamkeit komplexer gesundheitsfördernder Aktivitäten in komplexen sozialen Systemen oder Lebenswelten zu begreifen. … Gesundheitsförderungsmaßnahmen werden idealtypisch erst im Setting entwickelt bzw. werden bekannte Maßnahmen dort adaptiert und mit anderen Maßnahmen bedarfs- und bedürfnisgerecht kombiniert. Abhängige und mögliche unabhängige Variablen sind daher von vornherein nicht bekannt und setting- bzw. kontextspezifisch höchst variant.« (Elkeles & Broesskamp-Stone 2012). Es ist aber davon auszugehen, dass Maßnahmen, die wie Tabakrauchentwöhnung, Bewegungsprogramme etc., die in RCT gut untersucht und als im engen Sinne Evidenz basiert gelten (vgl. Sockoll et al. 2008), auch nach BZGA-Verständnis weiterhin zu den evidenzbasierten Interventionen in der Gesundheitsförderung zählen. Eine Abgrenzung von einfachen und komplexen Interventionen ist damit definitorisch erneut nicht möglich.

Zusammenfassend lassen die o. g. Begriffsverständnisse erkennen, dass komplexe Interventionen in erster Linie im Hinblick auf deren Kontextsensitivität problematisiert werden. Mit der Idee der komplexen Systeme, wie sie in der Annäherung an komplexe Intervention in der Gesundheitsförderung aufscheint, richtet sich der Blick aber auch auf ein anderes, nämlich systemtheoretisches Verständnis von Komplexität.

3.2 Komplexe Interventionen und komplexe Systeme

Das Verständnis komplexer Interventionen des MRC und die hierauf aufbauenden Definitionen weisen über den Geltungsanspruch des einfachen Wirkungsmodells der Abbildung 1 nicht hinaus. Dessen konstitutive Elemente erfahren zwar Komplikationen dadurch, dass sie sich nun durch eine Vielzahl von Komponenten, Akteuren oder Interventionsebenen operationalisieren. Die Intervention wird aber weiterhin als aus separaten Elementen bestehend aufgefasst. Das Verständnis ist zudem widersprüchlich, als dass eine Wirkungserwartung formuliert wird, nach der zwar alle Elemente a priori als essentiell betrachtet werden, trotzdem aber hierunter Elemente definitorisch hervorgehoben werden, die als eigentliche Wirkungsträger aufzufassen sind. Auch die explizit zugelassenen Einflüsse durch die Zielsubjekte der Intervention ergeben sich erneut nur durch die Vielzahl und folglich in der Variabilität von Verhalten und Effekten. Diese gewählte Kompliziertheit der Intervention würde zwar durch eine entsprechend kompliziertere Mathematik zur Umsetzung des Wirkungsmodells führen, dessen erkenntnistheoretische Grundlage aber nicht verändern.

Im Gegensatz hierzu tritt nach dem oben angeführten Verständnis in der Gesundheitsförderung

ein gänzlich anderes Bestimmungselement komplexer Interventionen hinzu, nämlich die Interventionen in komplexe Systeme, also in Systeme mit interdependenten, nicht linear interagierenden Komponenten. Kernbegriff in der Theorie komplexer Systeme ist die Emergenz, also das Erscheinen von Effekten, die – obwohl Systemantwort – nicht aus den Einzelkomponenten des Systems vorhersagbar sind. Ohne den Bedarf mystifizierender Deutungen, handelt es sich bei Emergenz nach transdisziplinären wissenschaftlichen Verständnis um eine Folge von Rückkopplungsprozessen in nicht linearen Dynamiken, die seit langen gut untersucht und systematisiert wurden. Auch vollständig determinierte und mathematisch gut beschreibbare Prozesse können Emergenz zeigen und ihre Systemzustände unvorhersagbar machen. Dabei können Interventionen in komplexe Systeme immer auch einfache Interventionen einschließen, die zur Initiierung von Kausalketten führen können. Rickles (2009) verweist auf ein Beispiel, dass sich aus den theoretischen Modellen Thomas Schellings über Segregation ergibt. Hierin wird gezeigt, wie minimale Unterschiede in den Präferenzen von Bewohnern etwa bezüglich Hautfarbe und Einkommen ihrer Nachbarn eine anfänglich gut integrierte Gemeinschaft zu einer vollständigen Segregation führt (Schelling 1978). Ein präventionsnahes Beispiel also für einen unerwünschten, vermutlich intentionskonträren emergenten Effekt.

Der Emergenzbegriff beschreibt entgegen seiner weitläufigen Verwendung keineswegs nur »positive«, systemstabilisierende Effekte. Im Gegenteil, Emergenzforschung fokussiert auf die Vermeidung von unvorhersehbaren Effekten. Die Beobachtung eines Systems und das Erkennen der Anfangsbedingungen und Einflüsse, die ein Systemgleichwicht gefährden steht im Vordergrund. Die "normal accidents theory" von Charles Perrow (1984) etwa leitet aus den Komplexitätseigenschaften von gekoppelten Systemen ab, dass auch die seltensten Systemzustände, also etwa katastrophale Störungen von großtechnischen Anlagen, normal sind. Während also Komplexitätseigenschaften eines Systems einen Theorierahmen zur Vermeidung von Katastrophen liefern, scheint es kein Modell dafür zugeben, wie unter Ausnutzung von Emergenz eine gewünschte Änderung gezielt erreicht werden kann. Allgemeiner formuliert es Luhmann (1984): »Komplexität … heißt Selektionszwang, Selektionszwang heißt Kontingenz, Kontingenz heißt Risiko«.

Zusammenfassend kann also hervorgehoben werden, dass das Begriffsverständnis komplexer Intervention weder einheitlich noch trennscharf ist. Als gemeinsames Element tritt die Vielzahl von Interventionskomponenten, Akteuren oder Kontexten auf. Obwohl als Idee in den Definitionen zur Kennzeichnung der Besonderheit der Intervention enthalten, ist nicht erkennbar, dass komplexe Interventionen tatsächlich die Ausnutzung der Komplexitätseigenschaften adaptiver Systeme in Aussicht nehmen.

4 Herausforderungen an den Wirkungsnachweis durch komplexe Interventionen?

Die in der Diskussion hervorgehobenen Probleme, die sich aufgrund von komplexen Interventionen für die Evaluation ergeben, lassen sich drei Perspektiven zuordnen. Die Perspektive der Multiplizität scheint in den medizinischen Disziplinen eingenommen zu werden, während die Perspektive der Kontextsensitivität in der Gesundheitsförderung betont wird. Die Perspektive komplexer Systeme dient vor allem als Referenzpunkt der Andersartigkeit komplexer Interventionen.

Multiplizität

In der Medizin scheinen die Probleme komplexer Intervention in den damit verbundenen Schwierigkeiten bei der Zusammenschau der Ergebnisse gesehen zu werden: »Studien, die die Wirksamkeit eines Programms erstmals experimentell evaluieren, werden mit Studien, die das Programm in einen anderen Kontext übertragen, synthetisiert.« Mühlhauser et al. (2011). Mit dieser Sichtweise wiederholt sich in Deutschland eine Diskussion, die international bereits nach Erscheinen der ersten o. g. MRC-Publikation geführt wurde. Komplexe Interventionen stellen danach ein Problem dar, weil die Interventionen oft unzureichend beschrieben sind und sich somit die Heterogenität von Interventionseffekten nicht auf Merkmale der Intervention zurückführen lassen (vgl. Shepperd et al. 2009). Das Problem besteht also nicht für die Evaluation komplexer Interventionen, sondern für die Ergebnissynthese einer Anzahl solcher

Evaluationen in systematischen Reviews. Anders gewendet handelt es sich also um ein Problem für die Evidenzbasierung, nicht für die Evaluation. Aus dieser Perspektive ist keine besondere Beschränktheit des Methodenarsenals der Wirkungsevaluation zu erkennen.

Kontextsensitivität

In der Gesundheitsförderung wird das Problem der komplexen Interventionen ebenfalls in dem Kontext gesehen. Während sich die unterschiedlichen Kontextbedingungen in der Perspektive der Multiplizität aber etwa aus der Vielzahl der Akteure ergeben, gleichsam also Folge eines notwendig differenten Interventionsmerkmals oder der Teilpopulationen sind, geht das Verständnis in der Gesundheitsförderung (nach den BZGA-Leitbegriffen) weiter. Hiernach ist die Interventionslage so grundsätzlich einzigartig, dass nicht einmal über abhängige und unabhängige Variable a priori geurteilt werden kann. Übersetzt in das o. g. Wirkungsmodell heißt dies, dass einer Intervention ein solches eben nicht zu Grunde gelegen hat. Nach diesem Verständnis müssten Maßnahmen der Gesundheitsförderung als intentionslose Interventionen verstanden werden, da auf einen Zielparameter verzichtet wird, dessen Beeinflussungsabsicht die Intervention leitet. Die Evaluation solcher Maßnahmen müsste sich entsprechend mit der Beschreibung der Prozesse und Strukturen zufrieden geben, eine Wirkungsevaluation wäre nicht möglich.

Auch ohne diese Transzendenz von Ursache und Wirkung ist die Sichtweise, gesundheitsbezogenes Handeln als Intervention sui generis zu verstehen, durchaus auch in anderen Disziplinen wie etwa der Homöopathie vorzufinden (vgl. Bödeker 2003). Was heute als Gegenpol der Evidenz basierten Medizin gesehen wird, folgt einem Leitbild, das noch vor kurzem das der gesamten Medizin war und auch heute noch in standespolitischen Diskussionen gehört wird. Desrosieres (2005) beschreibt diesen Standpunkt: »Für die traditionell eingestellten Ärzte, …, ist Medizin eine Kunst, die auf der Intuition und dem Instinkt des Praktikers aufbaute: Intuition und Instinkt manifestieren sich im Verlauf des singulären Kolloquiums zwischen Arzt und Patient und führten zu einer Indikation, die aus der Individualität des betreffenden Falles resultierte. Jeder Versuch, diesen Fall mit einer generischen Kategorie zu vergleichen, würde die Spezifität dieser persönlichen Wechselwirkung und die auf Erfahrung begründete Intuition des Falls zerstören.« Diese Einzigartigkeit wird auch aktuell in der Psychoanalyse diskutiert: »Das Heilsame der Therapie wäre gerade nicht die Wirksamkeit, die sich in messbare Parameter gießen lässt. … Denn in der Psychotherapie geht die Behandlung gerade nicht darin auf, was getan wird, sondern die Güte einer Therapie bemisst sich auch und gerade danach, mit welcher persönlichen Einstellung und Motivation, mit welchem Geist sie vollzogen wird.« (Maio 2011). Bei dieser Betonung der Kontextabhängigkeit ist fraglich, ob mit der Berücksichtigung des Therapeuten als konfundierender Faktor im o. g. Wirkungsmodell, dem Verständnis der Intervention Genüge getan wäre. Vermutlich geht der Einwand tiefer und bezweifelt, dass eine Subjekt-Objekt-Trennung in der Intervention möglich ist. Aus der Perspektive der Einzigartigkeit komplexer Interventionen besteht damit keine Herausforderung für die Wirkungsevaluation, denn sie wäre schlicht nicht möglich.

Komplexität

Die Perspektive der komplexen Systeme und damit definitorisch eine deutliche Abgrenzung zu nur komplizierten Intervention wird in der Komplementär- und Alternativmedizin eingenommen. Walach und Pincus (2012) stellen etwa Ergebnisse eines Diskurses von Komplementär- und Alternativmedizinern sowie Systemtheoretikern zusammen. Zielsetzung des Diskurses war es “This allows us to adequately describe and address ‘What happens to the person because of ‚it‘ … before asking ‘Does ‘it’ work” (Verhoef et al. 2012). Damit wird die schon oben abgeleitete Betrachtungsweise betont: Nicht mehr die intentionsgemäße Wirkung der Intervention steht im Fokus der Beurteilung, sondern die Beobachtung des Interventionsprozesses selbst. Das macht Sinn, wenn man das Verständnis von komplexen adaptiven Systemen zu Grunde legt. Diese sind ja gerade wie oben hergeleitet durch Unvorhersagbares charakterisiert. Die Beobachtung kann Ordnungsstrukturen der Systeme identifizieren, die die Wahrscheinlichkeit eines Auslenkens vom Gleichgewichtszustand verringern. Damit stellen sich aber auch gänzlich andere Anforderungen an das Studiendesign und

die Studiendurchführung. Im Hinblick auf das Studiendesign ist die Wirkungsevaluation damit als reine Beobachtungsstudie angelegt und Attribuierungen von Effekten zu Komponenten der Intervention sind auf einem allenfalls niedrigen Evidenzniveau angesiedelt. Im Hinblick auf die Studiendurchführung müssten die Probanden über das Fehlen einer Wirkungsvorstellung und sogar auf die fehlende Intention, eine solche jemals zu generieren, aufgeklärt werden. Ansonsten wäre die ethische Basis dieser Studien zweifelhaft. So fordert etwa die World Medical Association in ihren ethischen Prinzipien: "In medical research involving competent human subjects, each potential subject must be adequately informed of the aims, methods, sources of funding, ... the anticipated benefits and potential risks of the study and the discomfort it may entail, and any other relevant aspects of the study."(WMA 2008).

Zusammenfassend kann hervorgehoben werden, dass auch aus der Perspektive der komplexen Intervention in komplexe Systeme keine Aussagen über die wirkungsbestimmende Relevanz der Einflussfaktoren abgeleitet werden können. Einerseits ist nämlich eine Folge der Intervention nicht bestimmten Faktoren attribuierbar und zweitens sind sie per definitionem nicht reproduzierbar. Eine Ableitung von Handlungsempfehlungen ist aufgrund von komplexen Interventionen nicht möglich, eine Herausforderung an die Wirkungsevaluation stellt sich somit nicht.

5 Umgehen mit komplexen Interventionen bei der Wirkungsevaluation

Wie ausgeführt, sind die gängigen Vorstellungen komplexer Interventionen keine paradigmatisch neuen Herausforderungen für die Evaluation von Wirkungen der Interventionen. Das Verständnis von komplexer Intervention als einzigartig im Interventionskontext sowie als Beeinflussung eines komplexen Systems schließt eine Wirkungsevaluation im Sinne der gängigen Kausalitätstheorien aus. Evaluation müsste sich hier auf die Beschreibung der Prozesse als Fallstudien beschränken. Das Verständnis komplexer Intervention als Problem für die Wissenssynthese ist nicht Gegenstand dieser Betrachtungen, denn Evaluation von Interventionen haben eigene Zielsetzungen und gehen einer Evidenzbasierung voraus. Empfehlungen, wie bei der Erstellung systematischer Reviews mit Heterogenität der einbezogenen Studien umgegangen werden kann, liefert etwa das Handbuch der Cochrane Collboration (2011) und methodische Weiterentwicklungen (z. B. Verbeek et al. 2012). Die verbleibenden Herausforderungen komplexer Interventionen lassen sich durch drei Strategien begegnen (vgl. Bödeker 2011).

5.1 Schärfen der Fragestellung

Die eindeutige Festlegung einer Fragestellung, der Intervention und der betrachteten Endpunkte gehört zu den elementaren Anforderungen der Studienplanung, die etwa in den GRADE-Leitlinien nachzulesen sind (Langer et al. 2012). An dem Beginn auch einer komplexen Intervention muss die Darlegung der Begründetheit der Intervention und der wegen dieser spezifischen Intervention zu erwartenden Effekte stehen. Als Wegweiser eignen sich die entsprechende Fragen des MRC (2009): "Questions to ask yourself include:

- Are you clear about what you are trying to do: what outcome you are aiming for, and how you will bring about change?
- Does your intervention have a coherent theoretical basis? Have you used this theory systematically to develop the intervention?
- Can you describe the intervention fully, so that it can be implemented properly for the purposes of your evaluation, and replicated by others?
- Does the existing evidence – ideally collated in a systematic review – suggest that it is likely to be effective or cost effective? Can it be implemented in a research setting, and is it likely to be widely implementable if the results are favourable?

If you are unclear about the answers to these questions, further development work is needed before you begin your evaluation."

Bei einer Vielzahl von Interventionen in der Prävention ist davon auszugehen, dass diese Fragen nicht vorab ausreichend beantwortet werden konnten (vgl. Wijk, Mathiassen 2011). Insbesondere die »komplexe« Komponente der Intervention dürfte sich oft normativ aus Pragmatik, Machbarkeitsabwägungen und Theoriemangel ergeben.

5.2 Reduktion der Kompliziertheit

Das eingangs beschriebene einfache Wirkungsmodell ist so einfach nicht, da alle Konstituenten mit einer Vielzahl von methodischen Variationen eine Vielzahl von Interventionsszenarien modellieren können. Auch Interaktionen der Interventionskomponenten, deren nicht lineare Verknüpfung zu den unter Beobachtung stehenden Effekten wie auch etwaige hierarchische Strukturierungen von Teilpopulationen lassen sich hierdurch abbilden. Damit ist es für die Vielzahl der gängigen als komplexe Intervention aufgefassten Maßnahmen geeignet.

Für die komplexen Interventionen steht damit grundsätzlich der Untersuchungsansatz kontrollierter Studien zur Verfügung. Die komplexe Intervention kann dabei in toto hinsichtlich integraler Endpunkte oder als Teilinterventionen hinsichtlich spezifischer Endpunkte evaluiert werden. Eine optimierende Studienstrategie ist dabei die Reduktion der Kompliziertheit, die sich bereits in der Planungsphase aus der Beantwortung der o. g. Fragen des MRC ergeben kann. Es ist daher auch nicht überraschend, dass eine große Zahl von Evaluationsstudien zu im obigen Sinne komplexen Interventionen in Prävention und Public Health vorliegen, die auf kontrollierten und sogar randomisierten Studiendesigns basieren. Übersichten liefern etwa die Datenbanken der Cochrane- und der Campbell-Collaboration. Als besondere Herausforderung – auch jenseits der gegenwärtigen Komplexitätsdiskussion – wird die Evaluation sogenannter "universal policies" (WHO 2011) gesehen, in denen ein Kontrollgruppenansatz nicht immer möglich ist, da Maßnahmen zeitgleich auf die gesamte Bevölkerung ausgerollt werden. Aber selbst hier bieten sich internationale Vergleiche oder Vorher-Nachher-Vergleiche an. Rütten und Gelius (2012) weisen zudem darauf hin, dass mit dem Nachweis einer Wirksamkeit nicht schon der Nachweis eines geeigneten politischen Programms gegeben ist: »Politik, die sich mit Interventionen beschäftigt, deren Effektivität durch epidemiologische Studien nachgewiesen wurde, wird mit effektiver Politik verwechselt.«

Universal Policies bestehen idealerweise aus Maßnahmen, die vorher unter kontrollierten Bedingungen erprobt wurden. Ein "Policy Cycle" kann sicher stellen, dass weder mit Verweis auf einen fehlenden Wirkungsnachweis politisches Handeln ausbleibt, noch dass letzteres sich der Prüfung der Zielerreichung entzieht (Bödeker 2011). Ausgangspunkt dieses Kreislaufs wären politisch-normative Entscheidungen, die etwa in Gesetzen, Programmatiken oder Präventionszielen festgelegt sind. Zur Umsetzung dieser Programmatiken sind Interventionsprojekte zu konzipieren und zu erproben, denn die politische Zielsetzung beinhaltet nicht zwingend bereits auch das Wissen, wie Änderungen nachhaltig herbeigeführt werden können. Der Projekterprobung in Einzelstudien folgt die Projektevaluation, deren Ziel es ist zu beurteilen, inwiefern die angestrebten Prozesse, Strukturen oder Effekte erreicht worden sind. Liegen eine ausreichende Anzahl evaluierter Projekte vor, kann mit einer systematischen, über den Einzelfall einer Intervention hinausgehenden kritischen Betrachtung der Interventionserfolge begonnen werden. Dieser Prozess der Evidenzbasierung geht somit über die Evaluation von Einzelprojekten hinaus und zielt auf die Erhöhung der Beurteilungssicherheit der Interventionseffekte. Als Ergebnis der Evidenzbasierung können schließlich die erfolgreichen Interventionen und Interventionsmodalitäten hervorgehoben werden und als Qualitätsstandards für die Projekte im nachfolgenden Routinehandeln der Präventionsakteure dienen. Idealerweise sollten durch die an den Qualitätsstandards orientierten und in ausreichender Quantität durchgeführten Projekte die ursprünglich avisierten politischen Ziele erreicht werden können. Die Beurteilung der Zielerreichung schließt daher den Kreislauf und eröffnet ggf. die Diskussion über andere Interventionsstrategien oder andere Programmatiken.

5.3 Verzicht auf Wirkungsevaluation

Die dritte Strategie beim Umgang mit komplexen Interventionen ist schließlich der Verzicht auf eine Wirkungsevaluation. Ist eine Reduktion der Komplexität nicht möglich, weil das »Interventionssystem« als ontologische Entität nur durch alle einzelnen Konstituenten gesamt beschreibbar sein soll oder weil die Intervention die Komplexeigenschaften eines adaptiven Systems nutzen soll, so ist eine Evaluation der Wirkung dieser Intervention nicht möglich.

Der Verzicht auf eine Wirkungsevaluation ist unter diesen Bedingungen kein Mangel, denn die Interventionssituation wurde ja gerade entsprechend konzipiert (vgl. Trojan 2006). Ein Problem wäre es nur, wenn trotzdem ein Beweis-Bedarf gesehen wird, der vor dem Hintergrund eines hierzu nicht passenden Kausalitätsmodells erbracht werden soll. Anders gewendet, wenn das Vertrauen auf die Eignung der Intervention nicht groß genug ist, um ohne den Nimbus eines Wirkungsnachweises auszukommen.[1] Geschichte und Gegenwart zeigen, dass Gesellschaften hierzu in der Lage sind, da Wissenschaft nicht die einzige Inspirationsquelle für politisches Handeln ist (vgl. Humphreys, Piot, 2012).

1 »Beide Male besteht das Ziel der Quantifizierung nicht darin, die eigene Überzeugung zu sichern, sondern die Zustimmung einer gemischten und zerstreuten Gemeinschaft zu erlangen. Denn das wissenschaftliche Gemeinwesen, das die Quantifizierung pflegt und hochhält, ist ein Kollektiv, dessen Mitglieder höchst verschieden sein können …« Lorraine Daston 2003. Wunder, Beweise und Tatsachen. Fischer. Frankfurt

Literatur

Baumgärtner M (2007) Probleme einer theoretischen Analyse der Kausalrelation. In: Bundesanstalt für Arbeitsschutz und Arbeitsmedizin (Hrsg) Kausales Schließen auf der Grundlage von Beobachtungsstudien. Dortmund/Berlin/Dresden, S 16–34

Bödeker W (2006) Evidenzbasierung in Gesundheitsförderung und Prävention – Der Wunsch nach Legitimation und das Problem der Nachweisstrenge. In: Bödeker W, Kreis J (Hrsg) Evidenzbasierung in Gesundheitsförderung und Prävention. Wirtschaftsverlag NW, Bremerhaven – auch veröffentlicht in Prävention extra 3/2007, S 1–7

Bödeker W (2011) Evidenzbasierung ohne Kontrollgruppen – Wie können effektive Maßnahmen der betrieblichen Prävention erkannt werden? Zentralblatt für Arbeitsmedizin, Arbeitsschutz und Ergonomie 61: 78–83

Cameron ID, Murray GR, Gillespie LD et al. (2010) Interventions for preventing falls in older people in nursing care facilities and hospitals. Cochrane Database of Systematic Reviews 2010, Issue 1. Art. No.: CD005465. DOI: 10.1002/14651858.CD005465.pub2.

Cochrane Colloboration (2011) Cochrane Handbook for Systematic Reviews of Interventions. Version 5.1.0. http://www.mrc-bsu.cam.ac.uk/cochrane/handbook/ (Stand: 27.07. 2012)

Cohen JT, Neumann PJ, Weinstein MC (2008) Does prevention care save money? Health economics and the presidential candidates. N Engl J Med 358 (7): 661–663

DNebM (2012) http://www.ebm-netzwerk.de/kongress/2012/entwicklung-durch-austausch.pdf (Stand: 25.07.2012)

Elkeles T, Broesskamp-Stone U (2012) Evidenzbasierte Gesundheitsförderung. Leitbegriffe der Gesundheitsförderung. BZgA www.leitbegriffe.bzga.de

Humphreys K, Piot P (2012) Scientific evidence alone is not sufficient basis for health policy. BMJ 344: e1316

Langer G, Meerpohl JJ, Perleth M et al. (2012) GRADE-Leitlinien: 2. Formulierung der Fragestellung und Entscheidung über wichtige Endpunkte. Z Evid Fortbild Qual Gesundh wesen (ZEFQ) 106: 369–376

Luhmann N (1984) Soziale Systeme. Suhrkamp, Frankfurt

Maio G (2011) Verstehen nach Schemata und Vorgaben? Zu den ethischen Grenzen einer Industrialisierung der Psychotherapie. Psychotherapeutenjournal 2-20121: 132–138

McInnes E, Jammali-Blasi A, Bell-Syer SEM et al. (2011) Support surfaces for pressure ulcer prevention. Cochrane Database of Systematic Reviews 2011 (4). Art. No.: CD001735. DOI: 10.1002/14651858.CD001735.pub4.

Medical Research Council (2000) A framework for development and evaluation of RCTs for complex interventions to improve health.

Medical Research Council (2008) Developing and evaluating complex interventions: new guidance. www.mrc.ac.uk/complexinterventionsguidance (Stand: 20.05.2012)

Mühlhauser I, Lenz M, Meyer G (2011) Entwicklung, Bewertung und Synthese von komplexen Interventionen – eine methodische Herausforderung. Z Evid Fortbild Qual Gesundh wesen (ZEFQ) 105: 751–761

Rickles D (2009) Causality in complex interventions. Med Health Care and Philos 12: 77–90
Rütten A, Gelius P (2012) Evidenzbasierte Politik und nachhaltiger Wissenstransfer: Eine Perspektive für die Gesundheitsförderung in Deutschland. Gesundheitswesen 74: 224–228
Schelling TC (1978) Micromotives and macrobehavior. WW Norton, New York
Shepperd S, Lewin S, Straus S et al. (2009) Can we systematically review studies that evaluate complex interventions? PLoS Med 6(8): doi 10.1371
Sockoll I, Kramer I, Bödeker W (2008) Wirksamkeit und Nutzen betrieblicher Gesundheitsförderung und Prävention. IGA Report 13
www.iga-info.de
Trojan A (2006) Zu Chancen und Grenzen der Evidenzbasierung komplexer sozialer Interventionen. In: Bödeker W, Kreis J (Hrsg) Evidenzbasierung in Gesundheitsförderung und Prävention. Wirtschaftsverlag NW, Bremerhaven, S 73–109
Verbeek J, Ruotsalainen J, Hoving JL (2012) Synthesizing study results in a systematic review. Scand J Work Environ Health 38: 282–290
Verhoef MV, Kithan M, Bekk IR et al. (2012) Whole complementary and alternative medical systems and complexity: Creating collaborative relationships. Forsch Komplementmed 19 (suppl): 3–6
Walach H, Pincus D (2012) Kissing Descartes Good Bye. Forsch Komplementmed 19 (suppl): 1–2
WHO (2011) How can the health equity impact of universal policies be evaluated? Ed. Milton B, Moonan M, Taylor-Robinson D et al.
Wijk K, Mathiassen SE (2011) Explicit and implicit theories of change when designing and implementing preventive ergonomics interventions – a systematic literature review. Work Environ Health 37 (5): 363–375
WMA (2008) Declaration of Helsinki – Ethical Principles for Medical Research Involving Human Subjects
http://www.wma.net/en/30publications/10policies/b3/ (Stand: 23.07.2012)

Entwicklung, Bewertung und Synthese von komplexen Interventionen – eine methodische Herausforderung[1]

Ingrid Mühlhauser, Matthias Lenz, Gabriele Meyer

1 Einleitung

Viele medizinische Maßnahmen sind komplexe Interventionen. Sie bestehen aus mehreren Einzelkomponenten, die sich wechselseitig bedingen und ihrerseits in komplexe Kontexte implementiert werden. Beispiele dafür sind Stroke Units, Disease Management Programme oder Projekte zur Verbesserung der Krankenhaushygiene. Ähnliche komplexe Interventionen gibt es in assoziierten Berufs- und Handlungsfeldern. Zum Beispiel Prävention von Sturz und Dekubitus in der Pflege, Ernährungs- und Sportprogramme in Schulen, Prävention posttraumatischer Störungen, Früherkennung von Kindesmisshandlung und -verwahrlosung, Verringerung von Jugendkriminalität, Prävention von Unfällen im Straßenverkehr oder Web-basierte Lernprogramme.

Die Evaluation von Einzelmaßnahmen, wie die Behandlung mit einem Medikament, in randomisierten kontrollierten Studien (RCT) und deren systematische Übersichtsarbeiten ist vergleichsweise einfach. Die Wirksamkeit und der Nutzen und Schaden von komplexen Interventionen hingegen ist sehr viel schwerer zu erschließen. Der Beitrag der Einzelkomponenten zum Gesamtergebnis und die Interaktionen mit dem Setting bleiben häufig unklar. Seit einigen Jahren werden darum differenzierte methodische Verfahren zur Entwicklung, Bewertung und Synthese von komplexen Interventionen diskutiert.

Ziel dieses Artikels ist es, den Unterschied zwischen Einzelinterventionen und komplexen Interventionen herauszuarbeiten sowie methodische Ansprüche an die Entwicklung, Bewertung und Synthese von komplexen Interventionen zur Diskussion zu stellen.

2 Methodische Leitfäden zur Entwicklung und Evaluation komplexer Interventionen

Im Juni 2011 wurde systematisch nach Methodenpapieren zur Entwicklung und Evaluation komplexer Interventionen recherchiert. Die Datenbanken PubMed, Embase und PsycINFO wurden unter Verwendung von Suchbegriffen wie "complex intervention*", "multifaceted intervention*" in Kombination mit "methods" durchsucht. Die vollständige Recherchestrategie ist auf Anfrage bei den Autoren erhältlich. Die Internetseiten der *Campbell Collaboration*, der *Cochrane Collaboration*, des Instituts für Qualität und Wirtschaftlichkeit im Gesundheitswesen (IQWiG), des Deutschen Instituts für Medizinische Dokumentation und Information (DIMDI), des britischen *Social Care Institute for Excellence* (SCIE), des britischen *Centre for Reviews and Dissemination* (CRD), des britischen *Medical Research Council* (UKMRC), des britischen *National Institute for Health and Clinical Excellence* (NICE), des *International Network of Agencies for Health Technology Assessment* (INAHTA), des US-amerikanischen *Institute of Medicine* (IOM) und der *Agency for Healthcare Research and Quality* (AHRQ) wurden gesichtet. Eingeschlossen wurden Methodenpapiere bzw. methodische Leitfäden zur Entwicklung und Evaluation komplexer Interventionen, die die methodischen Herausforderungen deren Komplexität übergreifend berücksichtigen. Ausgeschlossen wurden Publikationen, die sich speziell mit Teilmethodiken (z. B. Methodik kontrollierter Studien oder Fokusgruppen) beschäftigen, sowie Übersichtsarbeiten, die sich allgemein mit Forschungsmethoden auseinandersetzen.

Die systematische Datenbankrecherche erzielte 1.261 Treffer. Aus eigenen Publikationsbeständen und nach Screening der Internetseiten kamen sieben Publikationen hinzu. Nach Screening von Titeln und Abstracts wurden 1.225 Treffer ausgeschlossen und 43 Publikationen im Volltext begutachtet. Das Flussdiagramm nach PRISMA und die Ergebnistabellen der Recherche sind auf Anfrage bei den Autoren erhältlich. Davon entspra-

1 Nachdruck aus: Mühlhauser I et al. (2011) Z. Evid. Fortbild. Qual. Gesundh. wesen (ZEFQ) 105: 751–761

chen 38 Arbeiten nicht den Einschlusskriterien. Hierzu zählten neben forschungsmethodischen Übersichtsartikeln (Blackwood 2006; Campell et al. 2007; Egan et al. 2009; Glasziou et al. 2010; Lenz et al. 2007; Shepperd et al. 2009; Walach et al. 2006; Pfaff et al. 2009; Raspe 2009) Publikationen, die sich mit spezifischen methodischen Aspekten beschäftigen: 1) quantitative Methoden in Übersichtsarbeiten (Lancaster et al. 2010; Ogilvie et al. 2008), 2) qualitative neben quantitativen Methoden bzw. "Mixed Methods" (Lewin et al. 2009; Palinkas et al. 2011; Paterson et al. 2009; Schifferdecker, Reed 2009; Yoshikawa et al. 2008; Andrew, Halcomb 2006; Curry et al. 2009; Kelle, Krones 2010; Miller et al. 2003; Miller, Fredericks 2006; Nastasi, Hitchcock 2009; Pluye et al. 2009; Whitley 2007), 3) "Realist Review" (Hunt, Sridharan 2010; Pawson et al. 2005) und 4) Evaluation komplexer Interventionen im Rahmen von RCTs und Prozessevaluation (Oakley et al. 2006; Emsley 2010; Perera et al. 2007).

Fünf relevante methodische Leitfäden wurden identifiziert:

1) Cochrane *"Handbook for Systematic Reviews of Interventions"* (Higgins, Green 2011),
2) Leitlinien *"Developing and evaluating complex interventions"* des UKMRC (Craig et al. 2008),
3) *"Guidance for undertaking reviews in health care"* des CRD (Centre for Reviews and Dissimination 2009),
4) *"Systematic research reviews"* des SCIE (Social Care Institute for Excellence 2010) und
5) *"Systematic Reviews in Social Sciences"* von Petticrew & Roberts (Petticrew, Roberts 2006).

Die fünf Leitfäden überschneiden sich inhaltlich. Eine strukturierte Anleitung zur Entwicklung komplexer Interventionen im Gesundheits- und Medizinbereich bietet vor allem die Leitlinie des UKMRC (Craig et al. 2008). Die Methodenpapiere des CRD (Centre for Reviews and Dissimination 2009) und des SCIE (Social Care Institut for Excellence 2010) verweisen auf diese Leitlinie. Das Cochrane Handbook (Higgins, Green 2011) richtet sich vor allem an Review-Autoren. Petticrew & Roberts diskutieren in ihrem als Buch erschienenen Leitfaden *"Systematic Reviews in Social Sciences"* (Petticrew, Roberts 2006) wesentliche Aspekte aus den anderen Methodenpapieren aus sozialwissenschaftlicher Perspektive. Mark Petticrew ist zudem Ko-Autor der Leitfäden des UKMRC, CRD und SCIE. Alle fünf Leitfäden sehen die Berücksichtigung qualitativer und quantitativer Methoden für die Entwicklung und Evaluation komplexer Interventionen in sehr ähnlicher Weise vor. Die zentralen methodischen Aspekte haben wir aus den fünf Leitfäden extrahiert und in einen konzeptionellen Rahmen integriert (siehe Abb. 1). Der konzeptionelle Rahmen soll die Zusammenhänge zwischen den wesentlichen methodischen Aspekten visualisieren. Er beinhaltet zusätzlich Referenzen zu methodischen Teilverfahren, z. B. CONSORT Statement (Boutron et al. 2008).

3 Sind nicht alle Interventionen komplex?

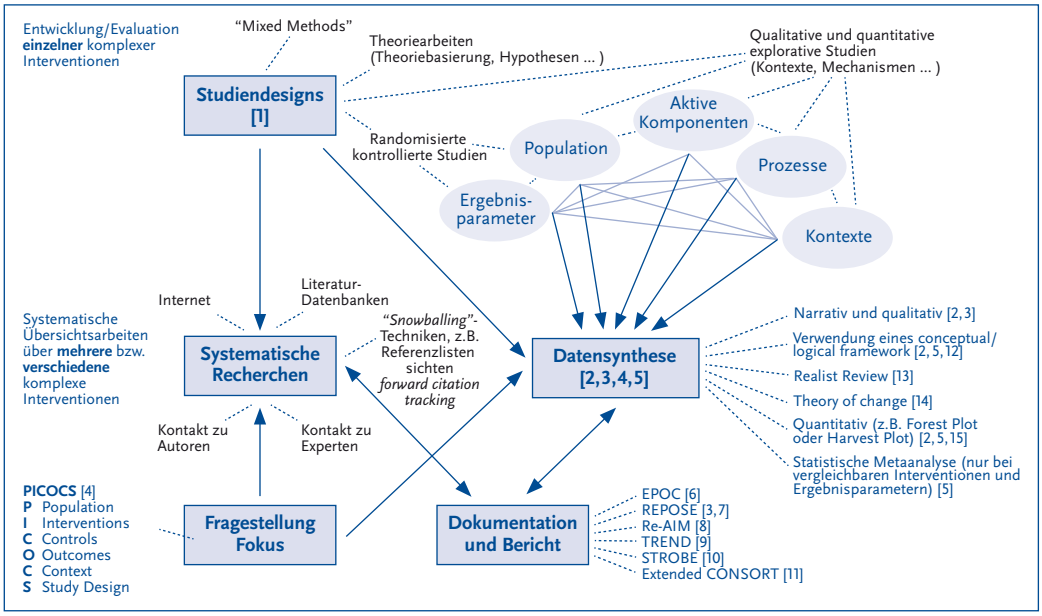
Ein Arzneimittel lässt sich als singuläre Intervention definieren. Mit der Zulassung durch die Arzneimittelbehörde sind die Entwicklungs- und Prüfphasen I bis III abgeschlossen, das Arzneimittel liegt somit in standardisierter Form vor. Die Wirksamkeit wird in RCTs durch verblindeten Vergleich mit einem Placebo-Präparat oder Standardmedikament nachgewiesen.

Es folgen kontrollierte Phase IV Studien zur Implementierung der neuen medizinischen Behandlung einschließlich Dokumentation der Sicherheit und Nutzen-Kosten Analysen.

Genau genommen besteht auch eine Arzneimittelbehandlung aus mehreren Komponenten. Die klinischen Effekte sind von verschiedenen Variablen abhängig, wie z. B. pharmazeutische Zubereitung, Applikationsform, Dosierung und Dauer der Behandlung, Patientenpopulation, spezifischem Krankheitsbild und Begleittherapie. Hinzu kommen Interaktionen mit anderen Arzneimitteln oder pathophysiologische Veränderungen wie eine Niereninsuffizienz, wie sie bei fast allen Medikamentenbehandlungen zu berücksichtigen sind.

Zudem kann auch in qualitativ hochwertigen RCTs die klinische Wirksamkeit eines Arzneimittels durch weitere Begleitfaktoren beeinflusst werden. So war z. B. in der Women's Health Initiative die postmenopausale Hormonbehandlung mit niedrigeren Blutzucker- und Cholesterinwerten assoziiert. Frauen in der Placebogruppe hatten folglich häufiger Statine erhalten, was zu einer Interaktion auf den primären kardiovaskulären

Abbildung 1
Vorschlag zu einem konzeptionellen Rahmen der Entwicklung und Evaluation kompexer Intervetionen
Quelle: Mühlhauser I et al. (2011) Z. Evid. Fortbild. Qual. Gesundh. wesen (ZEFQ) 105: 751–761



CONSORT = Consolidated Standards of Reporting Trials; EPOC = Effective Practice and Organisation of Care; RE-AIM = Reach, Efficacy, Adoption, Implementation, and Maintenance; REPOSE = Reporting of primary empirical research Studies in Education; STROBE = Strengthening the Reporting of Observational Studies in Epidemiology; TREND = Transparent Reporting of Evaluations with Nonrandomized Designs

(1) Craig et al. 2008, (2) Centre for Reviews and Dissimination 2009, (3) Higgins & Green 2011, (4) Petticrew & Roberts 2006, (5) SCIE 2010, (6) Cochrane Effective Practice and Organisation of Care (EPOC) Review Group 2000, (7) Newman & Elbourne 2005, (8) Glasgow et al. 2006, (9) Des Jarlais et al. 2004, (10) von Elm et al. 2008, (11) Boutron et al. 2008, (12) Briss et al. 2000, (13) Pawson et al. 2005, (14) Coote et al. 2004, (15) Ogilvie et al. 2008

Endpunkt führte (Writing Group for the Women's Health Initiative Investigators 2002). Medikamentenspezifische Beschwerden oder Komplikationen, wie das gehäufte Auftreten von gynäkologischen Blutungen unter Hormonbehandlung, können zu einer Intensivierung von Diagnostik und Versorgung führen. Auch die Effektstärke des Placebos bzw. der Standardtherapie beeinflusst das Ergebnis. Die in einem bestimmten RCT klinisch nachweisbare Wirksamkeit eines spezifischen Arzneimittels hängt somit vom Design der Studie, der methodischen Qualität der Studie, den Eigenschaften der Patientenpopulation und kontextuellen Faktoren ab. Die Bedingungen und Voraussetzungen bei der Evaluation von Medikamenten können jedoch weitestgehend standardisiert werden.

4 Wann sind komplexe Interventionen komplex?

Komplexe Interventionen bestehen aus mehreren interdependenten Komponenten. Ein Beispiel sind strukturierte Behandlungs- und Schulungsprogramme zur Insulintherapie von Patienten mit Diabetes Typ 1 (Mühlhauser, Berger 2002). Auch hier gibt es Arzneimittel-spezifische Faktoren. So hängt die Blutzuckerwirksamkeit von der Pharmakokinetik des Insulinpräparats ab, ob die Insuline variabel zu mischen sind und wann sie wie oft, mit welchem Spritz-Ess-Abstand, in welcher Dosierung verabreicht werden. Andere Einflussfaktoren sind jedoch von vergleichsweise größerer Bedeutung. Da die Insuline von den Patienten selbst appliziert und dosiert werden, wird die Wirksamkeit und Sicherheit der Insulinbehandlung entscheidend von den Möglichkeiten und Fähigkeiten des Patienten bestimmt, diese erfolgreich durchzuführen.

Zur Komponente Arzneimittel kommen also Komponenten wie Therapieschemata, Patientenschulung, Qualifikation und Motivation des Schulungs- und Behandlungsteams und Gesundheitssystembedingungen. Interdependenzen ergeben sich, da die Interventionskomponenten voneinander abhängen. Krankheits- und behandlungsbezogenes Wissen ist unverzichtbar, es reicht jedoch nicht aus. Die vermittelten Inhalte müssen Evidenz-basiert und handlungsrelevant sein. Blutzucker-/Stoffwechselselbstkontrollen sind unverzichtbar, jedoch kein Selbstzweck. Nur wenn die Patienten die Messungen korrekt durchführen, die Werte interpretieren können und durch adäquate Therapieanpassung reagieren, sind sie sinnvoll und nützlich. Eine Flexibilisierung des Tagesablaufs einschließlich variabler Nahrungszufuhr mit dem Ziel einer guten Blutzuckereinstellung ist nur durch eine angemessene, durch Patienten selbst gesteuerte Insulindosierung möglich. Ob die Therapie tatsächlich erfolgreich ist, hängt davon ab, ob und wie Stoffwechselselbstkontrollen durchgeführt werden, ob die Patienten passende (wirksame und sichere) Dosierungsregeln erhalten haben bzw. sich selbst erschließen, ob die Materialien ausreichend verfügbar sind und ob die Patienten die Motivation, die Kompetenzen und das Selbstvertrauen besitzen, die Therapien korrekt durchzuführen. Die Einstellungen und Haltungen des Schulungspersonals bzw. Behandlungsteams und die Voraussetzungen des jeweiligen Gesundheitssystems sind grundlegende Determinanten. Die Tabelle 1 illustriert die Komponenten, die den Erfolg eines Diabetes-Schulungsprogramms zur Verbesserung der Blutzuckereinstellung bestimmen können.

Einerseits kann bereits eine einzelne Komponente einer solchen multimodalen Intervention

Tabelle 1
Komponenten, die den Erfolg eines Diabetes-Schulungsprogramms zur Verbesserung der Blutzuckereinstellung bestimmen können – ein Beispiel

Komponenten	Inhaltliche Determinanten
Definition Patientenschulung *Zugrundeliegende Theorie/n?*	• Was verstehen die Autoren darunter? • Was verstehen die Studiennutzer darunter?
Ziele der Patientenschulung *Therapieziele?*	• Compliance vs. Selbstmanagement? • Arzt- vs. Patienten-definiert?
Inhalte Patientenschulung	• Evidenz-basiert? • Relevant? • Vollständig?
Komponenten Patientenschulung	• Stoffwechselselbstkontrolle – welche, wie? • Strikte Diät vs. freie Diät? • Selbstanpassung der Therapie durch Patienten? • Therapieverfahren?
Qualifikation Patientenschulung	• Wer schult? • Ausbildung der Schulenden? • Motivation, Einstellung und Überzeugungen des Behandlungsteams?
Organisation Patientenschulung	• Einzelschulung vs. Gruppenschulung • Ambulant – stationär? • Wie viele Unterrichtseinheiten?
Bedingungen Patientenschulung	• Gesundheitssystem? • Was ist verfügbar? • Was wird bezahlt?
Definition »Verbesserung der Blutzuckereinstellung«	• Individualisierte Therapieziele? • Intensität der medikamentösen Behandlung? • Liberalisierung von Diät und Lebensstil?
Heterogenität der Patientengruppen	• Erstschulung nach Diagnose • Wiederholungsschulungen • Bei Spätschäden

zum Misserfolg eines Programms führen. Andererseits sind es typischerweise mehrere unverzichtbare Komponenten, deren Zusammenwirken für den Erfolg einer komplexen Intervention verantwortlich sind.

Im Unterschied zur Entwicklung und Evaluation eines definierten Arzneimittels, macht die Multidimensionalität der komplexen Intervention eine sehr viel umfassendere Beschreibung und Begründung der übrigen Komponenten erforderlich, die die Gesamtheit der Intervention ausmachen. Während die Daten zur Zulassung eines Arzneimittels systematisch erhoben und dokumentiert werden müssen, fehlt bei komplexen Interventionen zumeist die ausreichend detaillierte und strukturierte Informationen zu ihrer theoretischen Fundierung und Entwicklung (Lenz et al. 2007; Shepperd et al. 2009; Warsi et al. 2004).

5 Komplexe Endpunkte

Ergebnisvariablen können eher singuläre, kombinierte oder auch komplexe Endpunkte sein.

Ein Beispiel für einen singulären Endpunkt ist die Gesamtmortalität. Ein häufig benutzter kombinierter Endpunkt ist »nicht-tödlicher Herzinfarkt, Schlaganfall oder Tod durch koronare Herzkrankheit« (Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen 2011). Im Gegensatz zum kombinierten Endpunkt ist ein komplexer Endpunkt direkt von anderen Erfolgsparametern abhängig. Ein Beispiel ist der HbA1c-Wert, wenn er zur Wirksamkeitsbewertung von strukturierten Diabetes-Behandlungs- und Schulungsprogrammen benutzt wird. Der HbA1c-Wert macht nur in der Zusammenschau mit anderen Erfolgsparametern Sinn. So kann bei Patienten mit Diabetes Typ 2 ein bestimmter HbA1c-Zielwert unterschiedlich erreicht werden, z. B. mit einer Intensivierung (The Action to Control Cardiovascular Risk in Diabetes Study Group 2008) oder auch Reduzierung der medikamentösen Behandlung (Kronsbein et al. 1988), je nachdem ob die komplexe Intervention eine gleichzeitige Gewichtsreduzierung vorsieht oder nicht. Bei Patienten mit Diabetes Typ 1 kann eine Verbesserung der HbA1c-Werte mit einer Zunahme oder Abnahme an schweren Unterzuckerungen einhergehen (Bott et al. 1997). Auch kann die Intervention eine Reduzierung der Arztkontakte aber auch Intensivierung der Betreuung vorsehen (Bott et al. 1997). Der HbA1c-Wert kann sowohl ein zwischen Patient und Arzt individuell definiertes Therapieziel implizieren oder aber ein für alle Patienten pauschal festgelegtes Therapieziel. Ein individuell festgelegter HbA1c-Zielwert kann z. B. bei fortgeschrittenem Alter oder Begleiterkrankungen durchaus höher liegen als der in Leitlinien als optimal definierte Grenzwert. Der komplexe Endpunkt HbA1c-Wert kann für ein Individuum also nicht interpretiert werden, ohne diese Begleitaspekte zu berücksichtigen.

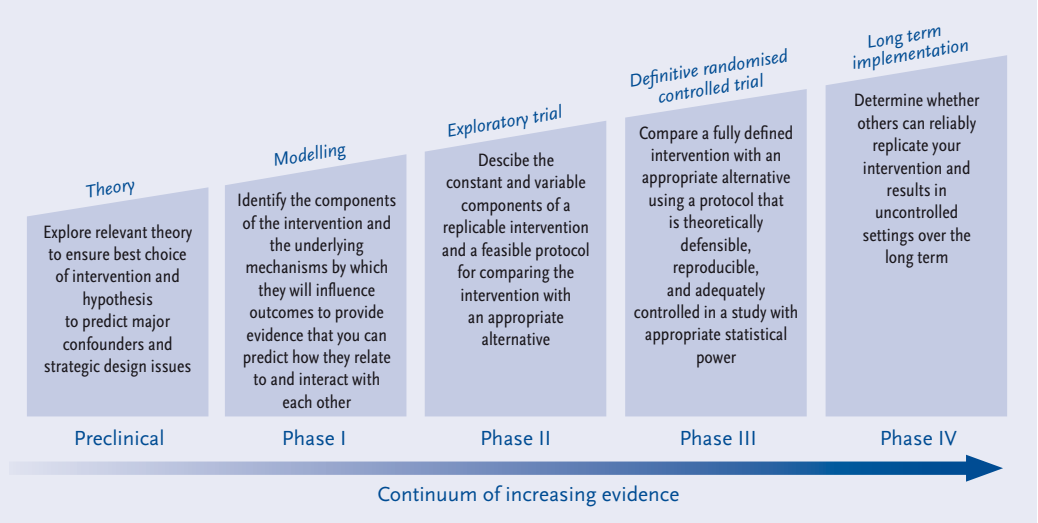
Auf der Methodenebene systematischer Übersichtsarbeiten dürften die Einzelkomponenten von komplexen Endpunkten nicht – wie sonst üblich – separiert voneinander betrachtet werden. Auch das GRADE System (Atkins et al. 2004), das sich inzwischen nicht nur für die Leitlinienerstellung, sondern auch für die Anfertigung von systematischen Reviews und Health Technology Assessments etabliert hat, bietet bislang keine methodische Lösung wie mit komplexen Endpunkten zu verfahren ist.

6 Entwicklung und Evaluation von komplexen Interventionen

Für die Entwicklung und Evaluation von komplexen Interventionen liegen seit 1999 Vorschläge vom britischen Medical Research Council (Campell et al. 2000; Craig et al. 2008) vor. Analog zur Entwicklung eines Arzneimittels wurden klinische Phasen I bis IV definiert (Abb. 2).

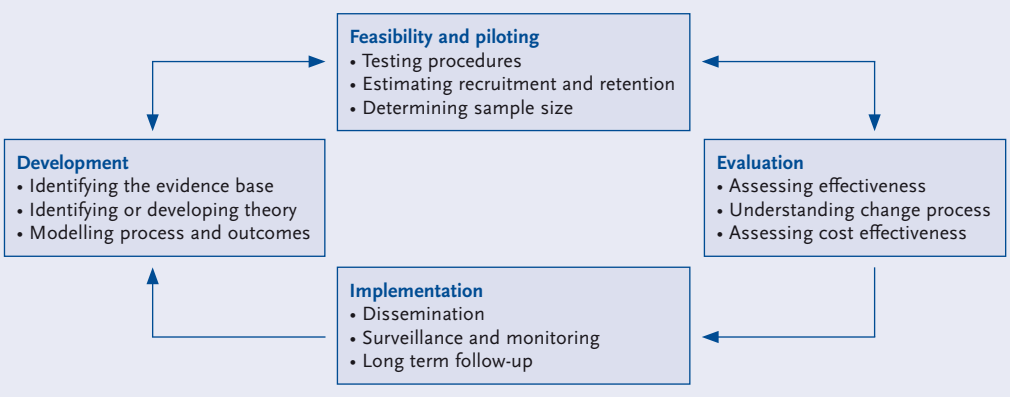
Die präklinische oder auch theoretische Phase dient zunächst der Identifizierung der bereits vorliegenden Evidenz, der Entwicklung erster Konzepte und der Bildung von Hypothesen über die intendierte Wirksamkeit der komplexen Intervention. Die Phase I beinhaltet den Entwurf und die Beschreibung einzelner Komponenten der komplexen Intervention, der Prototypenentwicklung und ersten Exploration von Wirkmechanismen und Erfolgsparametern. Die Phase II besteht in der Evaluation der Machbarkeit. Ermittelt werden beispielsweise unterschiedliche Implementierungsbedingungen, die Stabilität der komplexen Intervention in unterschiedlichen Kontexten und Effekte durch Modifikationen einzelner Komponenten in variierenden Kontexten. Ziel ist es

Abbildung 2
Sequenzielle Phasen der Entwicklung und Evaluation komplexer Interventionen
Quelle: Campbell M et al. BMJ 2000, 321: 694–696 (reproduziert mit Genehmigung der BMJ Publishing Group, 8/2012)



auch, die Bedingungen für ein RCT zu explorieren (z. B. Wirkhypothesen, Schätzungen der Stichprobe, Kontrollintervention). In der Phase III wird die komplexe Intervention als Ganzes mit einer angemessenen Alternative im RCT verglichen. Die Phase IV besteht schließlich in der Evaluation der Langzeitimplementierung. Untersucht werden die Übertragbarkeit, Reproduzierbarkeit und Langzeitwirksamkeit unter realen Bedingungen. Die Phasen beinhalten unterschiedliche qualitative und quantitative Forschungsmethoden, bauen aufeinander auf und bilden so ein Kontinuum ansteigender Evidenz (*continuum of increasing evidence* im Folgenden *increasing evidence*) (Campell et al. 2000). Vor der Revision des Konzeptes im Jahr 2008 hat die MRC-Arbeitsgruppe die Kritik zum

Abbildung 3
Kernelemente der Entwicklung und Evaluation komplexer Interventionen
Quelle: Craig P et al. (2008) UK Medical Research Council (reproduziert mit Genehmigung der BMJ Publishing Group, 8/2012)



Modell und die aufgezeigten Limitierungen analysiert (Craig et al. 2008). So impliziert das Konstrukt der *increasing evidence* eine der Praxis wenig entsprechende Linearität im Entwicklungs- und Evaluationsprozesses. Auch wurde die Variabilität von Versorgungskontexten nicht ausreichend berücksichtigt. Der revidierte Ansatz (Craig et al. 2008) beinhaltet vier Entwicklungs- und Evaluationsphasen (Abb. 3), die präklinische Phase und die Phase I wurden zur ersten Phase zusammengefasst, mögliche Evaluationsmethoden der einzelnen Phasen wurden spezifiziert, die Zirkularität der Prozesse und Kontextfaktoren sind berücksichtigt.

Der Kasten 1 illustriert ein eigenes Beispiel der Entwicklung und Evaluation einer komplexen Intervention aus der Altenpflege.

Kasten 1
Entwicklung komplexer Interventionen gemäß Leitlinie des UKMRC – ein Beispiel

In einem sechsjährigen, vom BMBF geförderten Vorhaben im Rahmen des Pflegeforschungsverbundes Nord, haben wir versucht, eine nachhaltig wirksame komplexe Intervention zur Reduktion von freiheitsentziehenden Maßnahmen (FEM) in Pflegeheimen gemäß UKMRC-Zyklus zu entwickeln. Wir haben zunächst eine epidemiologische Studie (Meyer et al. 2009) durchgeführt zur Bestimmung der Häufigkeit der Anwendung von FEM und der FEM-Arten (Gurte, Tische am geriatrischen Stuhl, beidseitige Bettgitter etc.) sowie Surveys zur Exploration von Haltungen, Einstellungen und Meinungen Pflegender und Angehöriger von Pflegeheimbewohnern gegenüber FEM (Hamers et al. 2009; Haut et al. 2010). Es folgte als notwendige Voraussetzung zur Modellierung der Intervention die Anfertigung eines systematischen Übersichtsartikels (Möhler et al. 2011) zur Identifikation wirksamer Interventionskomponenten. In einem nächsten Schritt wurde eine evidenzbasierte Leitlinie samt Implementierungshilfen entwickelt (Köpke et al. 2008). In der folgenden explorativen Phase wurden die Schulungs- und Implementierungsmaterialien und -programme in Fokusgruppen mit Pflegenden und Bewohnervertretern auf ihre Verständlichkeit und Machbarkeit untersucht. Zur Prüfung der Machbarkeit und Akzeptanz der Interventionsstudie und der Erhebungsinstrumente erfolgte eine Pilotphase mit vier randomisiert zugeteilten Einrichtungen (Haut et al. 2009). Es schloss sich ein Cluster-RCT an mit 36 Alten- und Pflegeheimen und ca. 4.000 Bewohnern. Die Interventionsgruppe erhielt die komplexe Intervention bestehend aus (1) Leitlinien-gestütztem Schulungsprogramm (1,5 h Schulung für alle Pflegenden in der Einrichtung, sechs Stunden Seminar für je zwei FEM-Beauftragte pro Einrichtung), (2) strukturierter Begleitung der FEM-Beauftragen über drei Monate, (3) Bereitstellung von Leitlinie und Informationsmaterial, (4) Reminder wie Poster, Becher oder Stifte mit einem projekteigenen Emblem. Die Kontrollgruppe erhielt eine Standardinformation (Haut et al. 2009). Auf der Prozessevaluationsebene wurden a) die Schulungen in den Einrichtungen der Interventionsgruppe (n=58) mit Pflegenden (n=569) anhand einer Erhebung zu Wissen und Selbstwirksamkeit evaluiert. Die FEM-Beauftragen wurden am Ende der Studie nach sechs Monaten interviewt, um die Barrieren und begünstigende Faktoren zu explorieren sowie das Ausmaß der Implementierung und Nachhaltigkeit. Da die Intervention klinisch wirksam und sicher zu einer Reduktion von FEM geführt hat (Köpke et al. 2012), wurden zum Zwecke der Dissemination die Einrichtungen der Kontrollgruppe im Anschluss an die Studie geschult und angeleitet, sowie die Leitlinie samt aller Implementierungsmaterialien kostenlos im Internet zur Verfügung gestellt (http://www.leitlinie-fem.de). Eine Implementierungsstudie befindet sich in Vorbereitung. Die Prozessevaluation liefert wiederum wertvolle Information zur Weiterentwicklung der Intervention zur Überwindung von hinderlichen Barrieren der Umsetzung.

7 Limitierungen der üblichen Verfahren der Synthese von Evidenz aus komplexen Interventionen

Mit der Zulassung eines Arzneimittels ist dieses in seiner Zusammensetzung standardisiert und lässt sich als definierte singuläre Intervention in klinischen Studien einsetzen. Die klinischen Studien unterscheiden sich dann im Wesentlichen durch Studien-, Patienten- und Kontextfaktoren. Die Intervention selbst ist jedoch weitgehend identisch. Das ursprüngliche primäre Ziel systematischer Übersichtsarbeiten und deren Meta-Analysen ist es, die Evidenz aus den verfügbaren RCTs zu einer Medikamentenbehandlung bei einer bestimmten Indikation quantitativ zusammenzufassen. Damit sollten die Erkenntnisse aus den oft viel zu kleinen klinischen Studien genutzt werden und Ressourcen verantwortungsvoll eingesetzt werden. Wirksame Behandlungen sollten frühzeitiger erkannt werden, der zu erwartende Effekt abgeschätzt werden, unnötige weitere Studien vermieden werden. Patienten sollten somit möglichst rasch wirksame und sichere Therapien erhalten und andererseits nicht nochmals unterlegenen Therapien (einschließlich Placebo) in experimentellen Studien ausgesetzt werden.

Systematische Übersichtsarbeiten sind in den letzten Jahrzehnten zunehmend an die Stelle traditioneller, selektiver Reviews (*literature reviews*, *narrative reviews*) getreten. Sie haben den Umgang mit klinischen Studien erheblich beeinflusst und zur Ausdifferenzierung empirisch geprüfter methodischer Standards klinischer Studien geführt. Der direkte Einfluss von systematischen Übersichtsarbeiten auf die Verbesserung der Patientenergebnisse ist jedoch schwer abzuschätzen. Aus kritischen Analysen zum Verzicht auf systematische Übersichtsarbeiten lässt sich jedoch indirekt auf ihren Nutzen schließen (Antman et al. 1992; Lau et al. 1992).

Nicht nur für die praktische Versorgung sind systematische Übersichtsarbeiten unentbehrlich geworden. Sie sollten den Ausgangspunkt einer jeden klinischen Studie darstellen. Handlungsleitend ist die Frage, ob es sinnvoll und geboten ist, eine weitere Studie durchzuführen. Renommierte medizinische Journale fordern inzwischen den Nachweis einer systematischen Übersichtsarbeit (Clark, Horton 2010). Methodiker rufen nachdrücklich dazu auf, nicht nur zur Bedarfsbestimmung einer klinischen Studie, sondern auch danach zur Einbettung der eigenen Ergebnisse eine systematische Übersichtsarbeit anzufertigen (Clarke et al. 2010).

Systematische Übersichtsarbeiten und deren Meta-Analysen haben ihre Limitierungen, im Wesentlichen durch die Qualität und Vollständigkeit der Berichterstattung der verfügbaren Daten. Die methodischen Probleme werden zunehmend erkannt und verbesserte Verfahren diskutiert, z. B. zum Umgang mit nicht publizierten Daten oder mit der Heterogenität der Studien (Riley et al. 2011). Für Meta-Analysen von Medikamentenstudien wurde kürzlich gefordert, die Gesamtheit der Evidenz zu betrachten, zum Beispiel die Ergebnisse aus unterschiedlichen Indikationsgebieten (Ioannidis, Karassa 2010). Der Schwerpunkt des methodologischen Diskurses liegt aktuell auf der Verbesserung statistischer Verfahren.

Kaum im Diskurs beachtet ist hingegen, ob die etablierten Verfahren komplexen Interventionen gerecht werden, d. h. die ursprünglichen, oben dargelegten Ziele systematischer Übersichtsarbeit erfüllt werden können.

In einer methodischen Übersichtsarbeit haben wir zu illustrieren versucht, dass systematische Übersichtsarbeiten derzeit nicht geeignet sind, komplexe Interventionen kritisch zu bewerten (Lenz et al. 2007). Über systematische Recherchen (PubMed, Cumulative Index to Nursing and Allied Health (CINAHL), Cochrane Library und HTA-Datenbanken) wurden zunächst systematische Reviews über Schulungs- und Selbstmanagementprogramme im Bereich Diabetes und Hypertonie identifiziert. Die der Ko-Autorin IM bekannte öffentlich verfügbare kumulative Evidenz von drei Schulungs- und Selbstmanagementprogrammen wurde als Referenzstandard benutzt, um zu untersuchen, ob die Programme und die dazugehörigen verfügbaren Publikationen angemessen berücksichtigt und bewertet wurden.

Vierzehn systematische Reviews wurden analysiert, darunter drei Cochrane Reviews. Zwölf der 14 Reviews bezeichneten die Programme korrekterweise als »multidimensional« oder »komplex«. Die Ergebnisse unserer Analysen zeigten dennoch erhebliche Defizite. In Reviews mit vergleichbaren Fragestellungen wurden verschiedene Publikationen identischer Referenzprogramme analysiert.

Identische Programme wurden zwischen verschiedenen aber auch innerhalb derselben Reviews unterschiedlich klassifiziert. In sechs Reviews wurden mittels meta-analytischer Verfahren Einflüsse von Einzelkomponenten (z. B. Dauer der Schulung) über unterschiedliche Schulungsprogramme hinweg analysiert. Deren klinische Heterogenität (z. B. unterschiedliche zugrunde liegende pädagogische Ansätze und Studiensettings) wurde nicht berücksichtigt. Interdependenzen zwischen Komponenten und Kontexten wurden nicht evaluiert. Auch die komplexen Interdependenzen zwischen komplexen Endpunkten (z. B. HbA1c und Hypoglykämien) und Kontexten und Zielen der Schulungsprogramme wurden nicht berücksichtigt.

Auf Basis unserer Ergebnisse haben wir schließlich sechs Kernpunkte zu einer angemessenen Evaluation komplexer Interventionen formuliert:

1) Alle Publikationen zur Entwicklung, Evaluation und Implementierung einer komplexen Intervention sollten berücksichtigt werden.
2) Die Literaturrecherchen sollten entsprechend angepasst werden (keine Limitierungen auf Studiendesign, spezifische Zielgruppen und Publikationszeiträume; "Snowballing"-Techniken wie Referenzlisten sichten, *forward citation tracking*, systematischer Autorenkontakt).
3) Die aktiven Komponenten sollten beschrieben und begutachtet werden.
4) Komplexe Interventionen sollten nicht Kategorien zugeordnet werden, die sich auf abhängige Komponenten beziehen.
5) Alle patientenrelevanten Endpunkte sollten angemessen berücksichtigt werden.
6) Das Poolen von Endpunkten über verschiedene komplexe Interventionen ist wegen inhärenter klinischer Heterogenität meist inadäquat.

Stattdessen sollten gegenseitige Abhängigkeiten zwischen Ergebnisparametern, Prozessen und Zielen analysiert und berichtet werden.

8 "Further research is needed"

Häufig lautet die Schlussfolgerung einer systematischen Übersichtsarbeit *"further research is needed"*. Dieser Appell erscheint unlogisch, wenn mehrere Dutzend Studien bereits vorliegen, die sich durch klinische und statistische Heterogenität auszeichnen. Am Beispiel Sturzprävention sei dies illustriert. Sturzprävention für Senioren besteht oft aus Risikoeinschätzung und Angebot Risikoadaptierter Interventionen. Die häufig als multifaktorielle Interventionen bezeichneten Maßnahmenpakete können neben Schulung und Information, Wohnraumanpassung und körperlichem Training, Medikamentenreview, Anpassung der Sehhilfen oder der Schuhe, Angebot von Hilfsmitteln oder Hüftprotektoren beinhalten. Ein Cochrane Review zum Thema (Cameron et al. 2010; Gillespie 2009) wurde kürzlich Setting-spezifisch in Sturzprävention im stationären Umfeld, d. h. Alten- und Pflegeheim und Krankenhaus (Cameron 2010), und Sturzprävention in der Häuslichkeit (Gillespie et al. 2009) unterteilt. Die Anzahl der eingeschlossenen RCTs ist beeindruckend: n = 41 Studien mit 25.422 Teilnehmern sowie n = 111 Studien mit 55.303 Teilnehmern. Die Schlussfolgerungen sind hingegen eher vage. Gillespie et al. (2009) resümieren aufgrund der klinischen und statistischen Heterogenität der eingeschlossenen Studien, dass die Kontexte in denen die vermutlich wirksamen Interventionen ihren klinischen Nutzen entfalten können, zu untersuchen bleiben. Auch Cameron et al. (2010) konstatieren ausgesprochene Inkonsistenz der Studienergebnisse und weiteren Forschungsbedarf.

Für die Bestimmung der aktiven Komponenten innerhalb komplexer Interventionen werden statistische Verfahren, z. B. Metaregressionsanalysen (Bower et al. 2006) diskutiert. Diese post hoc-analytischen Verfahren können zweifelsohne kein Ersatz sein für eine sorgfältige prospektive Entwicklung und Exploration der komplexen Intervention und der Interdependenzen mit kontextuellen Faktoren. Sie haben Observationscharakter und generieren maximal Hypothesen.

9 Sind andere Formen der Evidenz-Synthese besser geeignet für komplexe Interventionen?

Naheliegend ist es, sich anstatt der (ausschließlichen) Meta-Analyse explanatorischen Formen der Synthese zuzuwenden. Ein Modell zur rigorosen und strukturierten narrativen Synthese von Daten aus RCTs für die Fälle systematischer Reviews, in denen eine Meta-Analyse nicht möglich oder empfehlenswert ist, wurde kürzlich vorgelegt (Rodgers et al. 2009)

In der Pflegewissenschaft und Public Health erfreut sich die Realist Evaluation bzw. der Realist Review seit einigen Jahren großer Beliebtheit (Kane et al. 2010; Rycroft-Malone et al. 2010; Walshe, Luker 2010). Daher gilt es, diese Form der Evidenzsynthese hier eingehender zu diskutieren. Vorgeschlagen wird auch die Kombination von Realist Review und systematischem Review nach Cochrane oder Campbell Methoden (Van der Knaap et al. 2008).

Beim Realist Review stehen nicht die Fragen *"Does it work"* oder *"What works"* im Mittelpunkt, sondern die Frage *"What works for whom in what circumstances"*. Der Realist Review (Pawson et al. 2005) möchte diese komplexe Frage beantworten und unterscheidet sich nicht nur in seinem Fokus, sondern auch in seinen Methoden von einem konventionellen systematischen Review. Realist Reviews verlaufen weniger linear und sequentiell als konventionelle Reviews. Anfänglich definierte Fragen führen zu mehr Fragen und daher auch zur Auswertung diverser Informationsquellen mit verschiedenen Studiendesigns in dem Versuch zu erklären, warum, wann und wie die Intervention funktioniert. Ein Realist Review beginnt mit der einschlägigen Lektüre zur Identifikation der Faktoren mit dem größten vermeintlichen erklärendem Wert. Ein provisorisches explanatorisches Modell inkludiert die Faktoren und Behauptungen und ist handlungsleitend für den folgenden Review (Leeman et al. 2010). Der Realist Review geht davon aus, dass der Wirkungsweise einer komplexen Intervention ein Modell zugrundliegt, das erklärt, wie ein messbarer Effekt verursacht wird. Der entscheidende Aspekt ist der Kontext, in dem das Programm bzw. die komplexe Intervention stattfindet. **C**(ontext) – **M**(echanism) – **O**(utcome) Konfigurationen werden untersucht. Die identifizierte Literatur wird auf das provisorische explanatorische Modell angewendet und verfeinert dieses iterativ oder modifiziert es. Am Ende des Prozesses steht ein revidiertes explanatorisches Modell (Leeman et al. 2010).

Selbstredend ist ein Realist Review eine post hoc Synthese von Daten aus unterschiedlichen Quellen, auch inkonsistenter interner Validität. Ein demgemäß abgeleitetes Erklärungsmodell zur Wirksamkeit kann unserer Meinung nach nicht bedeuten, dass die Implementierung der in Frage stehenden komplexen Intervention nicht unter kontrollierten Bedingungen erfolgen soll.

Die Ergebnisse eines Realist Reviews können bei hohem Arbeitsaufwand erstaunlich marginal ausfallen. Dies sei an einem Beispiel illustriert: Wong et al. (2010) intendieren, Theorie-basierte Kriterien für die Entwicklung und Evaluation Web-basierter medizinischer Ausbildungskurse zu generieren. Nach umfangreichen Literaturrecherchen werden aus 12.586 Treffern schließlich 249 Studien jedweden Designs eingeschlossen und anhand des aufwändigen Realist Evaluation Verfahrens analysiert. Zwei Theorien, *Davis's Technology Acceptance Model* and *Laurillard's model of interactive dialogue*, werden identifiziert, die die Variation in den Ergebnissen und in der Zufriedenheit mit dem Angebot erklären. Das Ergebnis des Realist Reviews ist wenig spektakulär. Erstens sollen Web-basierte Kurse ihre Zielgruppe aktiv mit einbinden. Demnach steigt die Akzeptanz Web-basierter medizinischer Ausbildungskurse, wenn die Lernenden einen Vorteil empfinden gegenüber Nicht-Internet-basierten Angeboten, die technische Benutzung einfach und das Angebot kompatibel mit den eigenen Werten und Normen ist. Zweitens wird Interaktion hoch geschätzt. Lernende möchten mit einem Tutor oder anderen Lernenden in einen Dialog treten und ein formatives Feedback erhalten. Wong et al. (2010) reklamieren in der Diskussion ihres Artikels die limitierte Güte der ausgewerteten Studien. Wie jede andere Sekundäranalyse, kann die Güte der Synthese nur so gut sein wie die ihr zugrunde liegenden Daten. Nach Lektüre des Realist Reviews drängt sich dem Leser die Frage auf, ob nicht ein geringerer Aufwand zu vergleichbaren Ergebnissen geführt hätte. Beispielsweise durch strukturierte Befragung von Experten, die Web-basierte Kurse entwickelt haben.

10 Schlussfolgerung

Die prospektive Entwicklung und Evaluation komplexer Interventionen gemäß der UKMRC Leitlinie sowie die Bereitstellung aller Ergebnisse aus Pilotphasen und Begleitevaluationen kann nicht durch post hoc Ansätze ersetzt werden. Die Zusammenfassung der Forschung zu einer sorgfältig und langwierig entwickelten und erfolgreich evaluierten komplexen Intervention mit detaillierter Beschreibung der theoretischen Grundlagen und Bedingungskomponenten, der Erfolgs-befördernden und hinderlichen Faktoren stellt ebenso eine Form der Evidenz-Synthese dar. Sie ist für den Nutzer und Entscheidungsfinder im Gesundheitswesen relevanter und nachvollziehbarer als die grobe und teilweise irreführende Schlussfolgerung aus einer gemäß den üblichen Methoden angefertigten, nicht das gesamte Kontinuum der Evidenz sichtenden systematischen Übersichtsarbeit oder aus einer post hoc Synthese von Evidenz aus diversen Quellen unterschiedlicher Qualität wie im Realist Review.

Die Gesamtheit der Evidenz ist unverzichtbar, wenn darüber entschieden werden soll, ob eine Intervention in ein anderes Gesundheitssystem übertragbar ist. Herkömmliche systematische Übersichtsarbeiten im Interventionsbereich fokussieren auf RCTs. Deren Stellenwert soll hier keinesfalls in Abrede gestellt werden. Auf dem Kontinuum der Evaluation komplexer Interventionen erfolgt ein RCT jedoch erst nach einschlägigen Vorarbeiten in einem weit gediehenen Stadium der Interventionsentwicklung und -evaluation (Abb. 2 und 3). Die Gesamtheit der Evidenz für eine komplexe Intervention ist in systematischen Übersichtsarbeiten, die den standardisierten Methoden folgen, nicht zu beurteilen, da nicht danach recherchiert wird und die einzelnen Publikationen oft auch mit den üblichen Suchstrategien in den aktuellen Datenbanken nicht identifizierbar sind (Lenz et al. 2007). Insofern liefern systematische Übersichtsarbeiten zu komplexen Interventionen unvollständige Informationen. RCTs zu komplexen Interventionen, die sorgfältig und langwierig vorbereitet sind, werden mit unzureichend entwickelten komplexen Interventionen kombiniert. Studien, die die Wirksamkeit eines Programms erstmals experimentell evaluieren, werden mit Studien, die das Programm in einen anderen Kontext übertragen, synthetisiert (Lenz et al. 2007).

Es drängt sich die Frage auf, ob systematische Übersichtsarbeiten ihrer ursprünglichen Intention bei komplexen Interventionen noch gerecht werden können und tatsächlich zu einem Mehrwert an Wissen führen, oder aber eher zum Selbstzweck werden.

Diese und ähnliche Fragen sind bisher, gemessen an ihrer Bedeutung, nicht hinlänglich international kritisch gewürdigt.

Literatur

Andrew S, Halcomb EJ (2006) Mixed methods research is an effective method of enquiry for community health research. Contemp Nurse 23: 145–153

Antman EM, Lau J, Kupelnick B et al. (1992) A comparison of results of meta-analyses of randomized control trials and recommendations of clinical experts. Treatments for myocardial infarction. JAMA 268: 240–248

Atkins D, Best D, Briss P et al. (2004) GRADE Working Group. Grading quality of evidence and strength of recommendations. BMJ 328: 1490

Blackwood B (2006) Methodological issues in evaluating complex healthcare interventions. J Adv Nurs 54: 612–622

Bott S, Bott U, Berger M et al. (1997) Intensified insulin therapy and the risk of severe hypoglycaemia. Diabetologia 40: 926–932

Boutron I, Moher D, Altman DG et al. (2008) Extending the CONSORT statement to randomized trials of nonpharmacologic treatment: explanation and elaboration. Ann Intern Med 148: 295–309

Bower P, Gilbody S, Richards D et al. (2006) Collaborative care for depression in primary care. Making sense of a complex intervention: systematic review and meta-regression. Br J Psychiatry 89: 484–493

Briss PA, Zaza S, Pappaioanou M et al. (2000) Developing an evidence-based guide to community preventive services – methods. The Task Force on Community Preventive Services. Am J Prev Med 18 (Suppl 1): 35–43

Cameron ID, Murray GR, Gillespie LD et al. (2010) Interventions for preventing falls in older people in nursing care facilities and hospitals. Cochrane Database Syst Rev 2010 (1): CD005465

Campbell M, Fitzpatrick R, Haines A et al. (2000) Framework for design and evaluation of complex interventions to improve health. BMJ 321: 694–696

Campbell NC, Murray E, Darbyshire J et al. (2007) Designing and evaluating complex interventions to improve health care. BMJ 334: 455–459

Centre for Reviews and Dissimination (2009) Systematic Reviews: CRD's guidance for undertaking reviews in health care
www.york.ac.uk/inst/crd/pdf/Systematic_Reviews.pdf (Stand: 16.06.2011)

Clark S, Horton R (2010) Putting research into context – revisited. Lancet 376: 10–21
Clarke M, Hopewell S, Chalmers I (2010) Clinical trials should begin and end with systematic reviews of relevant evidence: 12 years and waiting. Lancet 376: 20–21
Cochrane Effective Practice and Organisation of Care (EPOC) Review Group (2000) Data collection checklist. Ottawa
Coote A, Allen J, Woodhead D (2004) Finding out what works: understanding complex, community-based initiatives. London: King's Fund
Craig P, Dieppe P, Macintyre S et al. (2008) Developing and evaluating complex interventions: new guidance. UK Medical Research Council
www.mrc.ac.uk/complexinterventionsguidance (Stand: 16.06.2011)
Curry LA, Nembhard IM, Bradley EH (2009) Qualitative and mixed methods provide unique contributions to outcomes research. Circulation 119: 1442–1452
Des Jarlais DC, Lyles C, Crepaz N (2004) Improving the reporting quality of nonrandomized evaluations of behavioral and public health interventions: the TREND statement. Am J Public Health 94: 361–366
Egan M, Bambra C, Petticrew M et al. (2009) Reviewing evidence on complex social interventions: appraising implementation in systematic reviews of the health effects of organisational-level workplace interventions. J Epidemiol Community Health 63: 4–11
Emsley R, Dunn G, White IR (2010) Mediation and moderation of treatment effects in randomised controlled trials of complex interventions. Stat Methods Med Res 19: 237–270
Gillespie LD, Robertson MC, Gillespie WJ et al. (2009) Interventions for preventing falls in older people living in the community. Cochrane Database Syst Rev 2009 (2): CD007146
Glasgow RE, Klesges LM, Dzewaltowski DA et al. (2006) Evaluating the impact of health promotion programs: using the RE-AIM framework to form summary measures for decision making involving complex issues. Health Educ Res 21: 688–694
Glasziou P, Chalmers I, Altman DG et al. (2010) Taking healthcare interventions from trial to practice. BMJ 341: c3852
Hamers JPH, Meyer G, Köpke S et al. (2009) Attitudes of Dutch, German and Swiss nursing staff towards physical restraint use in nursing home residents, a cross-sectional study. Int J Nurs Stud 46: 248–255
Haut A, Köpke S, Gerlach A et al. (2009) Evaluation of an evidence-based guidance on the reduction of physical restraints in nursing homes: a cluster-randomised controlled trial. BMC Geriatrics 9: 42
Haut A, Köpke S, Gerlach A et al. (2010) Attitudes of relatives of nursing home residents towards physical restraints. J Nurs Scholarsh 42: 448–456
Higgins JPT, Green S (2011) Cochrane Handbook for Systematic Reviews of Interventions Version 5.1.0 (Stand: March 2011)
www.cochrane-handbook.org/ (Stand: 15.06.2011)
Hunt KS, Sridharan S (2010) A realist evaluation approach to unpacking the impacts of the sentencing guidelines. AJE 31: 463–485
Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen (IQWiG) (2011) Allgemeine Methoden. Entwurf für Version 4.0 vom 09.03.2011
www.iqwig.de/download/IQWiG_Entwurf_Methoden_Version_4-0.pdf (Stand: 01.08.2011)
Ioannidis JP, Karassa FB (2010) The need to consider the wider agenda in systematic reviews and meta-analyses: breadth, timing, and depth of the evidence. BMJ 341: c4875
Kane SS, Gerretsen B, Scherpbier R et al. (2010) A realist synthesis of randomised control trials involving use of community health workers for delivering child health interventions in low and middle income countries. BMC Health Serv Res 10: 286
Kelle U, Krones T (2010) "Evidence based Medicine" und "Mixed Methods" – wie methodologische Diskussionen in der Medizin und den Sozialwissenschaften voneinander profitieren könnten. Z Evid Fortbild Qual Gesundh wesen: 104: 630–635
Köpke S, Meyer G, Haut A et al. (2008) Methodenpapier zur Entwicklung einer Praxisleitlinie zur Vermeidung von freiheitseinschränkenden Maßnahmen in der beruflichen Altenpflege. Z Evid Fortbild Qual Gesundh wesen 102: 45–53
Köpke S, Mühlhauser I, Gerlach A et al. (2012) Effect of a Guideline-based Multi-Component Intervention on Use of Physical Restraints in Nursing Homes. A Cluster Randomized Controlled Trial. JAMA 307: 2177–2184
Kronsbein P, Jörgens V, Mühlhauser I et al. (1988) Evaluation of a structured treatment and teaching programme on non-insulin-dependent diabetes. Lancet 2(8625): 1407–1411
Lancaster GA, Campbell MJ, Eldridge S et al. (2010) Trials in primary care: statistical issues in the design, conduct and evaluation of complex interventions. Stat Methods Med Res 19: 349–377
Lau J, Antman EM, Jimenez-Silva J et al. (1992) Cumulative meta-analysis of therapeutic trials for myocardial infarction. N Engl J Med 327: 248–254
Leeman J, Chang YK, Lee EJ et al. (2010) Implementation of antiretroviral therapy adherence interventions: a realist synthesis of evidence. J Adv Nurs 66: 1915–1930
Lenz M, Steckelberg A, Richter B et al. (2007) Meta-analysis does not allow appraisal of complex interventions in diabetes and hypertension self-management: a methodological review. Diabetologia 50: 1375–1383
Lewin S, Glenton C, Oxman AD (2009) Use of qualitative methods alongside randomised controlled trials of complex healthcare interventions: methodological study. BMJ 339: b3496
Meyer G, Köpke S, Haastert B et al. (2009) Restraint use among nursing home residents: cross-sectional study and prospective cohort study. J Clin Nurs 18: 981–990
Miller CL, Druss BG, Rohrbaugh RM (2003) Using qualitative methods to distill the active ingredients of a multifaceted intervention. Psychiatric Services 54: 568–571
Miller SI, Fredericks M (2006) Mixed-methods and evaluation research: trends and issues. Qual Health Res 16: 567–579

Möhler R, Richter T, Köpke S et al. (2011) Interventions for preventing and reducing the use of physical restraints in long-term geriatric care. Cochrane Database Syst Rev 2: CD007546
Mühlhauser I, Berger M (2002) Patient education – evaluation of a complex intervention. Diabetologia 45: 1723–1733
Mühlhauser1 I, Lenz M, Meyer G (2011) Z Evid Fortbild Qual Gesundh wesen (ZEFQ) 105: 751–761
Nastasi BK, Hitchcock J (2009) Challenges of evaluating multilevel interventions. Am J Community Psychol 43: 360–376
Newman M, Elbourne D (2005) Improving the usability of educational research: Guidelines for the REPOrting of primary empirical research Studies in Education (The REPOSE Guidelines). Evaluation and Research in Education 18: 201–212
Oakley A, Stange V, Bonell C et al. (2006) Process evaluation in randomised controlled trials of complex interventions. BMJ 332: 413–416
Ogilvie D, Fayter D, Petticrew M et al. (2008) The harvest plot: a method for synthesising evidence about the differential effects of interventions. BMC Med Res Methodol 8: 8
Palinkas LA, Aarons GA, Horwitz S et al. (2011) Mixed method designs in implementation research. Adm Policy Ment Health 38: 44–53
Paterson C, Baarts C, Launso L et al. (2009) Evaluating complex health interventions: a critical analysis of the 'outcomes' concept. BMC Complement Altern Med 9: 18
Pawson R, Greenhalgh T, Harvey G et al. (2005) Realist review – a new method of systematic review designed for complex policy interventions. J Health Serv Res Policy 10 (Suppl 1): 21–34
Perera R, Heneghan C, Yudkin P (2007) Graphical method for depicting randomised trials of complex interventions. BMJ 334: 127–129
Petticrew M, Roberts H (2006) Systematic reviews in the social sciences: a practical guide. Malden: Blackwell
Pfaff H, Albert US, Bornemann R et al. (2009) Methoden für die organisationsbezogene Versorgungsforschung. Gesundheitswesen 71: 777–790
Pluye P, Gagnon MP, Griffiths F et al. (2009) A scoring system for appraising mixed methods research, and concomitantly appraising qualitative, quantitative and mixed methods primary studies in Mixed Studies Reviews. Int J Nurs Stud 46: 529–546
Raspe H (2009) Unterschiede in der Bewertung von medikamentösen und nichtmedikamentösen Maßnahmen? Z Evid Fortbild Qual Gesundhwes 103: 253–260
Riley RD, Higgins JP, Deeks JJ (2011) Interpretation of random effects meta-analyses. BMJ 342: d549
Rodgers M,Sowden A, Petticrew M et al. (2009) Testing methodological guidance on the conduct of narrative synthesis in systematic reviews. Evaluation 15: 047–071
Rycroft-Malone J, Fontenla M, Bick D et al. (2010) A realistic evaluation: the case of protocol-based care. Implement Sci 5: 38
Schifferdecker KE, Reed VA (2009) Using mixed methods research in medical education: basic guidelines for researchers. Med Educ 43: 637–644
Shepperd S, Lewin S, Straus S et al. (2009) Can we systematically review studies that evaluate complex interventions? PLoS Med 6: e10000086
Social Care Institute for Excellence (SCIE) (2010) SCIE systematic research reviews: guidelines (2nd edition, 2010) www.scie.org.uk/publications/researchresources/rr01.pdf (Stand: 16.06.2011)
The Action to Control Cardiovascular Risk in Diabetes Study Group (2008) Effects of intensive glucose lowering in type 2 diabetes. N Eng J Med 358: 2545–2559
Van der Knaap LM, Leeuw FL, Bogaerts S et al. (2008) Combining Campbell Standards and the realist evaluation approach: The best of two worlds? AJE 29: 48–57
von Elm E, Altman DG, Egger M et al. (2008) The strengthening the reporting of observational studies in epidemiology (STROBE) statement: guidelines for reporting observational studies. J Clin Epidemiol 61: 344–349
Walach H, Falkenberg T, Fonnebo V et al. (2006) Circular instead of hierarchical: methodological principles for the evaluation of complex interventions. BMC Med Res Methodol 6: 29
Walshe C, Luker KA (2010) 'District nurses' role in palliative care provision: a realist review. Int J Nurs Stud 47: 1167–1183
Warsi A, Wang PS, LaValley MP et al. (2004) Self-management education programs in chronic disease: a systematic review and methodological critique of the literature. Arch Intern Med 164: 1641–1649
Whitley R (2007) Mixed-methods studies. J Ment Health 16: 697–701
Wong G, Greenhalgh T, Pawson R (2010) Internet-based medical education: a realist review of what works, for whom and in what circumstances. BMC Med Educ 10: 12
Writing Group for the Women's Health Initiative Investigators (2002) Risks and Benefits of Estrogen plus progestin in healthy postmenopausal women. JAMA 288: 321–333
Yoshikawa H, Weisner TS, Kalil A et al. (2008) Mixing qualitative and quantitative research in developmental science: Uses and Methodological Choices. Dev Psychol 44: 344–354

Methodenprobleme bei der Evaluation komplexerer Sachverhalte: Das Beispiel Suchtprävention

Alfred Uhl

Es ist inzwischen zu einem »Dogma des Common-Sense« geworden, dass Prävention, wie auch alle anderen öffentlich finanzierten Aktivitäten, regelmäßig evaluiert werden sollte, um die Wirksamkeit der Maßnahmen nachzuweisen und so den Mitteleinsatz zu legitimieren. Um sich diesem Thema sinnvoll annähern zu können, ist es zunächst zweckmäßig zu klären, was »Evaluation« bedeutet und welche konkreten Optionen dem Evaluator offenstehen, bzw. welche Evaluationsstrategien an methodologischen, ethischen, ökonomischen und ontologischen Grenzen scheitern. Beide Themen sind äußerst komplex. Der Ausdruck »Evaluation« ist sehr vage und mehrdeutig – die Bedeutung reicht von »präzisem, experimentellem Wirkungsnachweis« bis hin zu »einen oder mehrere Aspekte im Zusammenhang mit dem Evaluationsgegenstand zu erfassen«. Die Entscheidung, was man tun darf bzw. tun muss, ist ethischer Natur und kann von der Forschung mittels Evaluation nicht beantwortet werden, auch wenn das immer wieder suggeriert wird.

Experimentelle Zugänge mit optimalen Kontrollbedingungen sind bei Evaluationen kaum herstellbar. Die wirkungsrelevanten Parameter zwischenmenschlicher Interaktion sind nur schwer identifizierbar und messbar. Mit konkreten Projekten erreichte Personenzahlen sind meist viel zu klein, um Effekte zufallskritisch absichern zu können. Kurzfristige Präventionsergebnisse sind wenig relevant, da es eigentlich um langfristige Effekte geht, aber Forschungsprojekte über viele Jahre werden selten durchgeführt, da man die Entscheidungsgrundlagen rasch benötigt. Hier könnte man noch viele weitere Probleme anführen.

Im folgenden Text wird abrissartig auf diese Probleme und einige weitere eingegangen, um einen Eindruck zu vermitteln, mit welchen Problemen Evaluatoren und Forscher konfrontiert sind, ohne dabei Anspruch auf Vollständigkeit zu erheben. Abschließend wird versucht, eine Perspektive zu entwickeln, wie man sich als Evaluator und Forscher in dieser schwierigen Situation verhalten sollte.

1 Was bedeutet Evaluation

1.1 Alltagsverständnis vs. professionelles Verständnis von »Evaluation«

Die meisten Menschen haben ein intuitives Verständnis darüber, was »Evaluation« bedeutet – nämlich die Beantwortung der Frage, wieweit ein bestimmter Evaluationsgegenstand[1] die an ihn gestellten Anforderungen erfüllt. Dieses intuitive Alltagsverständnis von »Evaluation« ist auch der Vorstellungswelt vieler Experten nicht fremd, greift aber zu kurz bzw. spiegelt nur einen möglichen Aspekt von dem wider, was im wissenschaftlichen und professionellen Sprachgebrauch unter »Evaluation« tatsächlich subsumiert wird. Im professionellen Sinn umfasst der Ausdruck »Evaluation« systematische Forschungsaktivitäten von der Planung und Entwicklung bis zur endgültigen Beurteilung eines Evaluationsgegenstandes. Die Diskrepanz zwischen dem Alltagsverständnis und dem professionellen Verständnis von »Evaluation« ist Quelle für gravierende Missverständnisse im Evaluationsdiskurs, die den Beteiligten aber oft gar nicht bewusst werden.

1.2 Vierdimensionales Klassifikationsschema der COST-A6-Expertengruppe

Zwischen 1994 und 1998 stand eine Arbeitsgruppe von 21 Personen aus 15 europäischen Staaten und Kanada im Rahmen des COST-A6-Programms der europäischen Kommission vor der Aufgabe, einen Überblick über evaluierte Suchtpräventionsprogramme in Europa zu präsentieren (Uhl 1998, 2000a). Rasch wurde klar, dass es wenig sinnvoll ist, in den beteiligten EU-Staaten nach evaluierten Primärpräventionskonzepten und -programmen zu suchen, wenn kein Konsens darüber besteht,

1 Im Zusammenhang mit Suchtprävention ist der Evaluationsgegenstand meist ein mehr oder weniger genau beschriebenes Konzept oder ein manualisiertes Programm.

was unter »Suchtprävention« und unter »Evaluation« überhaupt zu verstehen ist.

Beim ersten Treffen wurde rasch offensichtlich, dass fast jeder der Teilnehmer diese beiden Ausdrücke anders interpretierte. Daraufhin entschlossen sich die Teilnehmer, ein Subprojekt zur Präzisierung der Ausdrücke »Suchtprävention« und »Evaluation« zu beginnen. Als Forschungsstrategie wurde ein Delphi-Ansatz mit drei Durchgängen gewählt. Im Zuge dieses Prozesses ergaben sich die Evaluation betreffend vier Dimensionen (Datendimension, Zeitdimension, methodologische Dimension und Evaluatordimension), die von den Teilnehmern zu einem vierdimensionalen Klassifikationsschema integriert wurden.

1.2.1 Datendimension

Ein wichtiger Aspekt von »Evaluation« bezieht sich auf die Art der verwendeten Daten. Die COST-A6-Expertengruppe definierte, bezugnehmend auf Clayton and Cattarello (1991), die zwischen »Prozessevaluation«, »Ergebnisevaluation« und »Impactevaluation« unterschieden hatten, und auf Donabedian (1980), der in »strukturelle Qualität«, »Prozessqualität« und »Ergebnisqualität« unterteilte, fünf Datenarten, die in Zusammenhang mit Evaluation eine Rolle spielen:

- »Strukturelle Evaluation« baut auf Strukturdaten, die z. B. beschreiben, an wie vielen Orten und mit wie vielen Teilnehmenden eine Maßnahme umgesetzt wird, welche Kosten dafür anfallen, etc.
- »Kontextevaluation« basiert auf Kontextdaten, die Rahmenbedingungen erfassen, die einen Einfluss auf das Geschehen haben könnten, indem sie den Zusammenhang zwischen Interventionen und Auswirkungen moderieren. Eine gute Analyse des Kontexts ermöglicht Urteile darüber, wie und in welchem Ausmaß Evaluationsergebnisse auf andere Situationen übertragbar sind.
- »Ergebnisevaluation« basiert auf Ergebnisdaten; das sind jene Resultate, die über die Erreichung der definierten Interventionsziele Auskunft geben.
- »Impactevaluation« (dafür gibt es keine adäquate deutsche Übersetzung) basiert auf Impactdaten, die jene positiven und negativen Auswirkungen erfassen, die nicht als explizite Maßnahmenziele definiert waren bzw. mit denen man zunächst nicht gerechnet hatte.
- »Prozessevaluation« basiert auf Prozessdaten, die Hinweis darüber liefern, was im Laufe der Durchführung einer Maßnahme konkret passiert ist, um so relevante Abläufe zu dokumentieren und zu erklären. Dies umfasst Interventionen von Programmausführenden (Input), Reaktionen der Zielpersonen (Output) und relevante kontextuelle Bedingungen, die den Zusammenhang zwischen Input und Output beeinflussen. Ziel ist es, hypothetisch zu erklären, ob und warum Maßnahmen unter bestimmten Rahmenbedingungen bestimmte Resultate hervorrufen.

1.2.2 Zeitdimension oder Entwicklungsstadium des Evaluationsgegenstandes

Scriven (1967, 1991), einer der international bekanntesten Evaluationsexperten, entwickelte eine Unterteilung von »Evaluation«, die auf drei Entwicklungsstadien des Evaluationsgegenstandes Bezug nimmt. Diese dreistufige Unterteilung wurde von der COST-A6-Expertengruppe auf vier Stufen erweitert, indem die dritte Phase in eine Überprüfungsphase und eine Routinephase aufgeteilt wurde. Damit ergab sich die folgende Einteilung:

- »Präformative Evaluation« beschreibt jene Aktivitäten, die darauf abzielen, den vorläufigen Entwurf eines Evaluationsgegenstandes zu entwickeln. Es geht hier um einen rein gedanklichen Prozess ohne praktische, empirisch orientierte Schritte. Diese Konzeptphase wird als »präformative Phase« bezeichnet.
- »Formative Evaluation« beschreibt Aktivitäten, die aufbauend auf den in der präformativen Phase entwickelten vorläufigen Entwurf – durch wiederholte praktische Erprobung – den konkreten Evaluationsgegenstand formen. Formative Evaluation zielt auf die rasche und flexible Erfassung von Schwachstellen im Konzept mit dem Ziel, vorläufige Entwürfe so lange zu erproben und erfahrungsgestützt zu verbessern, bis sich ein Evaluationsgegenstand ohne offensichtliche Schwachstellen ergibt.

- »Summative Evaluation I« beschreibt Aktivitäten, um den nunmehr als fertig wahrgenommenen Evaluationsgegenstand zusammenfassend zu beurteilen. In dieser Überprüfungsphase, die man als »erste summative Phase« bezeichnen kann, geht es darum, sowohl erwartete als auch unerwartete Effekte des Evaluationsgegenstandes zu erfassen und zu dokumentieren, um ein zusammenfassendes Urteil über den Evaluationsgegenstand zu ermöglichen.
- »Summative Evaluation II« beschreibt Aktivitäten in der Routinephase. Diese beginnt, nachdem Erfolg angenommen wurde und der Evaluationsgegenstand zur routinemäßigen Anwendung gelangt. In dieser Phase, die man als »zweite summative Phase« bezeichnen kann, geht es darum zu gewährleisten, dass die Qualität der Programmdurchführung erhalten bleibt, und auch darum, nach unerwarteten längerfristigen Effekten bzw. relevanten Veränderungen in den Rahmenbedingungen Ausschau zu halten. Dieser Evaluationsansatz wird oft auch als »Qualitätssicherung« und/oder »Qualitätskontrolle« bezeichnet und stellt einen wichtigen Schritt dar, um die Anwendungstreue (Application Fidelity[2]) zu gewährleisten.

Identifiziert man in der ersten summativen Phase oder der Routinephase Probleme, so erfolgen in der Regel erneut formative Schritte, um den Evaluationsgegenstand zu optimieren. Bei gravierenden Problemen wird der Evaluationsgegenstand entweder endgültig aufgegeben oder es finden neuerlich ganz grundlegende Reflexionen, im Sinne einer präformativen Vorgangsweise, statt.

1.2.3 Methodologische Dimension

Ein dritter Zugang, um Evaluation zu klassifizieren, unterscheidet zwischen deskriptiver, explorativer and konfirmatorischer Evaluation. In Anlehnung an Tukey (1977) wird hier nach der Aussagekraft der evaluierenden Forschung unterschieden; d. h. danach, welche Schlüsse man erkenntnistheoretisch begründbar aus den Studienergebnissen ziehen darf.

- »Deskriptive Evaluation« steht für das bloße Erfassen und Dokumentieren von Phänomenen sowie für deren Kategorisierung und Zusammenfassung, ohne daraus neue Hypothesen ableiten zu wollen.
- »Explorative Evaluation« geht über die reine Deskription hinaus und formuliert Hypothesen. Explorative Forschung reicht dabei vom Erfassen grundlegender Informationen in eher wenig erforschten Wissenschaftsbereichen bis zur hypothesengeleiteten Entwicklung von neuen Modellen und Theorien. Es gibt keine strengen Regeln, welche Schritte bei exploratorischen Studien zulässig sind. Alles, was einen tieferen Einblick in relevante Zusammenhänge ermöglichen kann, ist zulässig – allerdings nur solange nicht behauptet oder suggeriert wird, dass endgültige Antworten und Beweise gewonnen werden konnten.
- »Konfirmatorische Evaluation« beschäftigt sich damit, präzise formulierte Hypothesen experimentell zu prüfen. Der »Goldstandard« der »konfirmatorischen Evaluation« sind randomisierte, kontrollierte Studien (RCTs), mit denen alleine sich Kausalität methodologisch fundiert belegen lässt. Aber RCTs sind aus technischen, ökonomischen, ethischen und epistemologischen Gründen im human- und sozialwissenschaftlichen Feld oft nicht realisierbar. In der Regel muss man hier mit weit weniger eindeutigen Forschungsansätzen vorlieb nehmen.

1.2.4 Evaluatordimension

Die Evaluatordimension bezieht sich auf die Position der Personen, die die Hauptverantwortung für die jeweiligen Evaluationen tragen. Man unterscheidet dabei zwischen:

- »interner Evaluation«, wo die Verantwortung für die Evaluation bei jenen Personen liegt, die das Konzept für den Evaluationsgegenstand entwickelt haben und/oder dieses routinemäßig anwenden, und
- »externer Evaluation«, wo die maßgeblichen Evaluationsschritte in den Händen von nicht direkt involvierten Evaluatoren liegen. Wie jene Personen, die Evaluationen planen, durchführen

2 "Application Fidelity" ist das Ausmaß, wie genau und originalgetreu eine Technik, eine Intervention oder ein Programm gemäß Manual, Protokoll oder Theorie eingesetzt wird.

und interpretieren zum Auftraggeber und zu den Projektverantwortlichen stehen, ist natürlich nicht unerheblich.

Auf die Vor- und Nachteile beider Zugänge wird in einem späteren Abschnitt noch genauer eingegangen werden.

1.2.5 DZME-Klassifikation

Die vier Dimensionen »**D**atendimension«, »**Z**eitdimension«, »**m**ethodologische Dimension« und »**E**valuatordimension« lassen sich zu einem vierdimensionalen Klassifikationsschema integrieren, das nach den Anfangsbuchstaben dieser Dimensionen »DZME-Klassifikation« genannt wurde und hier noch einmal zusammenfassend dargestellt wird:

- **D**atendimension
  - Strukturevaluation basierend auf Strukturdaten
  - Kontextevaluation basierend auf Kontextdaten
  - Ergebnisevaluation basierend auf Ergebnisdaten (explizit erwartete Auswirkungen)
  - Impactevaluation basierend auf Impactdaten (unerwartete Auswirkungen)
  - Prozessevaluation basierend auf Prozessdaten
- **Z**eitdimension (Entwicklungsstadium des Evaluationsgegenstandes)
  - Konzeptevaluation in der Konzeptphase (präformative Phase)
  - formative Evaluation in der Entwicklungsphase (formative Phase)
  - summative Evaluation I in der Erprobungsphase (erste summative Phase)
  - summative Evaluation II in der Routinephase (zweite summative Phase)
- **m**ethodologische Dimension
  - deskriptive Evaluation (Dokumentation – beschreibend)
  - explorative Evaluation (hypothesengenerierend)
  - konfirmatorische Evaluation (hypothesenprüfend)
- **E**valuatordimension
  - interne Evaluation
  - externe Evaluation

Der DZME-Ansatz zur Klassifikation von Evaluationsprojekten kann, sofern die Begriffe korrekt verstanden und verwendet werden, einen wichtigen Beitrag zur Präzisierung des Diskurses über Evaluation leisten. Der Ansatz reicht alleine aber nicht aus, um die Komplexität des Evaluationsbegriffes vollständig abzubilden. Mehrere wichtige Formen der Evaluation – das reicht von der Bedarfsprüfung (Need Assessment) bis zur Wirtschaftlichkeitsprüfung (Kosten-Nutzen oder Kosten-Effektivitätsanalysen) – werden dabei nicht erfasst. Es erschien der COST-A6-Arbeitsgruppe daher zweckmäßig, diesen eher abstrakten Klassifikationsansatz durch ein stärker inhaltlich orientiertes System zu ergänzen, wobei im inhaltlichen Klassifikationssystem auch einige Begriffe vorkommen, die im DZME-Klassifikationsansatz eine Rolle spielen (Uhl 1998, 2000a, 2000b).

1.3 »Evaluation« im Sinne der »Standards für Evaluation« der DeGEval

2002 stellte die deutsche Gesellschaft für Evaluation (DeGEval) eine Kommission aus sieben Evaluatoren und zwei Klientenvertretern zusammen, deren Aufgabe es war, einen Entwurf für deutsche Evaluationsstandards vorzulegen (Beywl 2003). Die von dieser Kommission erarbeiteten Standards und die einleitende Definition von »Evaluation« wurde in der Folge von der DeGEval angenommen.

Im Sinne der DeGEval Definition bedeutet »professionelle Evaluation«:

- Informationen sollen gesammelt und aggregiert werden, die es ermöglichen, ein Urteil über den Evaluationsgegenstand abzugeben – wobei die Evaluatoren nicht unbedingt selbst ein direktes Urteil über den Gegenstand abgegeben müssen. Manche Auftraggeber verwehren sich sogar nachdrücklich dagegen, dass von Ihnen beauftragten Evaluatoren wertende Schlüsse ziehen, sondern erwarten ausschließlich fundierte Grundlagen, um derartige Schlüsse ziehen zu können.
- Evaluationen sollten in den Händen von umfassend ausgebildeten, kompetenten Evaluatoren liegen.
- Evaluatoren sollten systematisch und empirisch vorgehen, was aber keinesfalls bedeutet, dass standardisierten Instrumenten der Vorzug zu geben ist.

- Evaluatoren sollten alle wesentlichen Entscheidungen und Maßnahmen bewusst setzen und alle Schritte so gut dokumentieren, dass diese für Dritte nachprüfbar sind.
- Es ist zwar grundsätzlich sinnvoll, sich bei der Implementierung neuer Strategien und/oder neuer Programme Gedanken über deren Evaluation zu machen bzw. eine solche mit zu planen, das bedeutet aber nicht, dass Evaluationsgegenstände, bei deren Implementierung keine Evaluation geplant wurde, nicht trotzdem sinnvoll evaluiert werden könnten.
- Oft wird unreflektiert mechanistisch gefordert, dass bei der Evaluationsplanung von Anfang an präzise Ergebniskriterien festzulegen seien, was häufig völlig unsinnig ist. Geeignete Zielkriterien kann nur formulieren, wer mit den Besonderheiten des Evaluationsgegenstands hinreichend vertraut ist. Es ist aber kaum realistisch, dass Evaluatoren bereits bei Angebotslegung, also vor dem verbindlichen Zuschlag, ausführliche Vorarbeiten leisten, um sich mit dem Evaluationsgegenstand im Detail auseinanderzusetzen. Wer als Evaluator mit offenen Augen durch den Evaluationsprozess geht, wird in der Regel laufend mit neuen Aspekten konfrontiert, die es zu erfassen und zu berücksichtigen gilt, und mit denen er ursprünglich nicht rechnen konnte. Sich ohne detaillierte Kenntnis des Forschungsgegenstandes ad hoc auf Zielkriterien festzulegen und diese starr zu messen, führt fast zwangsläufig dazu, dass wichtige Aspekte übersehen werden und irrelevanter Datenmüll gesammelt wird. Das fällt oft nur deswegen nicht auf, weil erfahrende Evaluatoren beliebige Ergebnisse frei assoziierend in einer Art und Weise interpretieren können, dass das eigentliche Scheitern der Evaluation den meisten Rezipienten des Berichts nicht bewusst wird. Evaluationsexperten, die das durchschauen, kritisieren solch eine Vorgangsweise in der Regel auch nicht öffentlich, weil sie bei ähnlich gearteten Projekten auch nicht anders vorgehen – also im Glashaus Sitzende nicht mit Steinen werfen wollen. Die DeGEval-Standards führen dazu Folgendes aus: »*Dabei ist zu bedenken, dass rigide Festlegungen bezüglich der Detailfragestellungen, der Methodik und des Vorgehens leicht zu Hemmschuhen werden können, die den Erkenntnisgewinn nachhaltig behindern*«.
- Gegen vorab festgelegte starre Zielkriterien spricht auch der Umstand, dass sich Personen in den zu evaluierenden Einrichtungen, um gut abzuschneiden, häufig an den vorgegebenen messbaren Kriterien orientieren und in der Folge von der Erfüllung der quantitativ schwer fassbaren, eigentlich relevanten Qualitätskriterien abrücken. Dazu kommt erschwerend, dass großer Dokumentationsaufwand zwangsläufig die Qualität der eigentlichen Arbeit verringert, sofern, wie dies oft der Fall ist, keine zusätzlichen Mittel zur Verfügung gestellt werden. Unter solchen Bedingungen kann das, was naive Auswerter aufzeichnen und als Qualitätssteigerung beschreiben, paradoxer Ausdruck massiven Qualitätsverlusts sein.

1.4 Pseudoevaluationen

Wenn ein kompetentes und motiviertes Team interessiert ist, Abläufe und Ergebnisse des Arbeitsfeldes zu optimieren, so ist der ideale Evaluationsansatz eine »formative, interne Evaluation« in enger Zusammenarbeit mit externen Evaluatoren. Das Team profitiert von den Erfahrungen der externen Evaluatoren und von deren Fähigkeit, neue Einsichten und Ideen aus der Außenperspektive einzubringen. Die externen Evaluatoren können inhaltlich auf die fachlichen Erfahrungen des Teams zurückgreifen. Diese Form der Evaluation minimiert Evaluationskosten, da externe Evaluatoren Aufgaben an interne Mitarbeiter delegieren, die diese im Rahmen ihrer Routinetätigkeit mit vergleichsweise geringem Zusatzaufwand erledigen. Darüber hinaus ist die Umsetzung der Evaluationsergebnisse garantiert, weil das Team im Zuge der formativen Evaluation neue Erkenntnisse sofort in die Abläufe integrieren kann und über das direkte Engagement im Prozess die gemeinsam erarbeiteten Resultate auch genau versteht.

Wenn allerdings vermutet wird, dass ein bestimmtes Programm oder eine bestimmte Institution bloß suboptimal funktioniert und das Team wenig Veränderungsmotivation zeigt, so sind kritische, externe Evaluationen zweckmäßig. In solchen Problemsituationen kommt es allerdings oft vor, dass maßgebliche Entscheidungsträger ihr Urteil – zu Recht oder zu Unrecht – bereits gefällt haben und eine externe Evaluation beauftragen, um

bereits getroffene Entscheidungen scheinbar »faktengestützt« legitimieren zu können. Evaluatoren, denen mehr oder weniger deutlich vermittelt wird, welche Ergebnisse vom Auftraggeber erwartet werden, können es sich kaum leisten, die Auftraggeber maßgeblich zu enttäuschen, ohne ihre berufliche Existenz nachhaltig zu gefährden. Derartige Evaluationen mutieren mit hoher Wahrscheinlichkeit zu Pseudoevaluationen. Die Gefahr, dass sich die Evaluierten gegen das abgekartete Spiel öffentlich wehren, zwingt die Evaluatoren allerdings auch hier zu einem gewissen Ausmaß an Objektivität und professionell vertretbarem Vorgehen.

Pseudoevaluationen passieren aber nicht nur dann, wenn Entscheidungsträger bereits gefällte Entscheidungen gegen die Weiterführung von Projekten, Abteilungen oder Institutionen legitimieren wollen, sondern auch dann, wenn sie bereits getätigte Ausgaben nachträglich rechtfertigen möchten. In diesem Fall decken sich die Interessen der Entscheidungsträger mit jenen der Projektverantwortlichen, die ja ebenfalls gut dastehen wollen. Ganz gleich, ob nun Entscheidungsträger oder Projektverantwortliche die Evaluationen beauftragen – ein deutlich negatives Ergebnis der Evaluation ist unwahrscheinlich, selbst wenn ein Projekt große Mängel aufweist.

1.5 Evalopathie

Pseudoevaluationen hat es immer gegeben, aber vieles ist früher nicht evaluiert worden. Seit einiger Zeit steigt jedoch der Druck, alles und jedes zu evaluieren. Wir sind mit einem Evaluationsboom konfrontiert und das Wort »Evaluation« ist in aller Munde. Das mag einerseits mit der angespannten wirtschaftlichen Situation zusammenhängen, in der Entscheidungsträger den Eindruck erwecken müssen, die begrenzten Mittel möglichst sparsam und optimal einzusetzen; andererseits unterliegen die Sozial- und Humanwissenschaften zunehmend dem von den Naturwissenschaften abgeleiteten Diktat, dass alles objektiv mess- und überprüfbar sein muss. So betont z. B. der Wiener Krankenanstaltenverbund (2012), bezugnehmend auf das seit den 1980er-Jahren populäre Schlagwort "New Public Management" (vgl. dazu Hood 1991), dass »*Qualität, Effizienz und Effektivität im Zentrum des Handelns stehen sollen*«. Erwähnt werden in diesem Zusammenhang Konzepte wie »*Controlling, Berichtswesen, Qualitätsmanagement, Informations- und Kommunikationsinstrumente sowie Leistungsvergleiche*«. In eine ähnliche Kerbe schlägt auch die immer populärer werdende Forderung nach einer »evidenzbasierten Politik« (siehe dazu Punkt 2).

All diese Schlagworte klingen zunächst unmittelbar überzeugend. Niemand wird ernsthaft bestreiten wollen, dass Qualität, Effizienz und Effektivität im Zentrum des Handelns stehen sollten, und natürlich ist es auch sinnvoll zu überprüfen, ob öffentliche Mittel zweckmäßig und sparsam eingesetzt werden. Aber wenn man das ohne Grundverständnis für die Grenzen des Messbaren plant, wenn man ohne Bezug zu konkreten Evaluationszielen umfassende Dokumentationsinstrumente konzipiert und implementiert, und wenn man vergisst zu beurteilen, ob der dafür notwendige Aufwand in einer vertretbaren Relation zum erwartbaren Nutzen steht, entwickeln sich sinnentleerte bürokratische Rituale, die jene Qualität torpedieren, zu deren Sicherung sie eingeführt wurden.

Die kritiklose Überzeugung, dass alles und jedes ständig dokumentiert und evaluiert werden muss, breitet sich aus wie eine ansteckende Krankheit, die man als »Evalopathie« bezeichnen könnte (Uhl 2000c). Da werden Daten produziert, die keinen echten Erkenntnisgewinn ermöglichen, aber Ressourcen verbrauchen, die weit nutzbringender eingesetzt werden könnten. Immer mehr Personen sind rund um die Uhr damit beschäftigt festzuhalten, was sie arbeiten würden, wenn sie nicht durch umfassende Dokumentationsverpflichtungen davon abgehalten würden. Die dermaßen »Kontrollierten« sitzen, wenn sie sich offen gegen die ausufernde Bürokratie wehren, meist auf dem kürzeren Ast. Widerstand erfolgt daher meist subtil. Statt sich offen quer zu legen, werden Pseudoevaluationen inszeniert und Datenmüll produziert. Die darauf aufbauenden Hochglanzberichte mit wenig aussagekräftigen Statistiken und Evaluationsergebnissen werden als unverzichtbar dargestellt.

Der große Erfolg von Begriffen wie »New Public Management«, »Qualitätssicherung«, »Evidenzbasiertheit« etc. lässt sich mit Liessmann (2005) über das Prinzip der »performativen Selbstimmunisierung« erklären. Derartige Schlagworte funktionieren nach Liessmann wie »Zauberwörter«. Liessmann schreibt dazu: »D*iese bezeichnen nie das, was die Wortbedeutung nahe legt, verbergen aber, was*

durch sie tatsächlich indiziert wird. Gelingen kann dieses Täuschungsmanöver nur, weil alle diese Begriffe dem Prinzip der performativen Selbstimmunisierung gehorchen. Wer Evaluation, Qualitätssicherung oder Internationalisierung sagt, hat immer schon gewonnen, da diese Begriffe ihre Negation nur um den Preis der Selbstbeschädigung zulassen. Denn natürlich will niemand in den Verdacht geraten, Leistungen nicht messen zu wollen, der Qualität kein Augenmerk zu schenken, sich vor dem Wettbewerb zu fürchten und in der Provinzialität zu versinken.«

Um nicht missverstanden zu werden: Ich wende mich hier nicht grundsätzlich gegen Evaluation – im Gegenteil, Evaluation erscheint mir wichtig. Es macht Sinn, unsere Arbeitsroutinen regelmäßig zu analysieren und darauf aufbauend zu verbessern – und all das kann man unter »Evaluation« subsumieren. Wir sollten aber unbedingt danach trachten, dass der Aufwand in einem ausgewogenen Verhältnis zum erwartbaren Nutzen steht und dass methodologische, ethische, ökonomische und ontologische Erkenntnisgrenzen nicht ignoriert werden.

Eine Evaluationspraxis jenseits von »Evalopathie« könnte sich durchaus entwickeln, wenn die Durchführbarkeitsstandards der DeGEval (Beywl 2003) verstanden und von der professionellen Welt auch ernst genommen würden. Dazu einige Zitate aus diesen Standards: »*Die Durchführbarkeitsstandards sollen sicherstellen, dass eine Evaluation realistisch, gut durchdacht, diplomatisch und kostenbewusst geplant und ausgeführt wird (...). Evaluationsverfahren, einschließlich der Verfahren zur Beschaffung notwendiger Informationen, sollen so gewählt werden, dass Belastungen des Evaluationsgegenstandes bzw. der Beteiligten und Betroffenen in einem angemessenen Verhältnis zum erwarteten Nutzen der Evaluation stehen (...). Die aus wissenschaftlicher Sicht aussagekräftigsten Methoden können oft nicht verwendet werden, da sie zu aufwändig (zeitraubend oder kostspielig) oder im entsprechenden Kontext ethisch nicht akzeptabel sind (...). Vor- und Nachteile sowie Aussagekraft der gewählten Verfahren sollen durch das Evaluationsteam transparent gemacht und begründet werden (...). Evaluationen sollen so geplant und durchgeführt werden, dass eine möglichst hohe Akzeptanz der verschiedenen Beteiligten und Betroffenen in Bezug auf Vorgehen und Ergebnisse der Evaluation erreicht werden kann (...). Der Aufwand für Evaluation soll in einem angemessenen Verhältnis zum Nutzen der Evaluation stehen (...). Gerade bei der Entscheidung über die Durchführung einer Evaluation sollen Kosten und Nutzen abgeschätzt werden.«*

2 Ethik und Suchtprävention: Welt-, Menschen- und Gesellschaftsbild

Es wird immer wieder gefordert, dass politische Entscheidungen sich auf wissenschaftliche Fakten gründen. Solche Entscheidungen können aber grundsätzlich nicht losgelöst von ethisch-weltanschaulichen Überzeugungen und den dahinter stehenden Welt-, Menschen- und Gesellschaftsbildern getroffen werden.

2.1 Evidenzbasierte Politik – z. B. im Bereich Suchtpolitik oder Alkoholpolitik

Auf den ersten Blick erscheint das Schlagwort »Evidenzbasiertheit«, das aus dem Bereich der Medizin kommt, auch in Verbindung mit Prävention oder Suchtpolitik durchaus sinnvoll. Wenn wir uns an der Konzeption von Sackett et al. (1996) orientieren, die in Zusammenhang mit evidenzbasierter Medizin »Evidenzbasiertheit« als »bewusste, explizite und vernünftige Verwendung der zum Zeitpunkt vorliegenden besten Evidenz« definierten, so ist dagegen kaum etwas zu sagen. Auch hier wirkt das zuvor beschriebene Liessmannsche Prinzip der »performativen Selbstimmunisierung« spontan gegen potenzielle Kritiker. Wie könnte man das scheinbare Gegenteil, die unreflektierte, implizite und unvernünftige Bezugnahme auf die schlechteste Evidenz vertreten, ohne sich lächerlich zu machen.

Geht man allerdings etwas in die Tiefe, so verliert der Ausdruck rasch an Glanz. Dieser Zugang ist keinesfalls so neu, wie suggeriert wird, und auch keinesfalls so faktenorientiert, wie der Ausdruck nahelegt. Viele Wissenschaftler der Vergangenheit waren auch schon der Meinung, optimale Schlüsse aus der besten vorhandenen Evidenz (= ihr persönlicher Kenntnisstand) zu ziehen. Das explizit so zu vertreten, wird als Überheblichkeit ausgelegt, aber wer sich nunmehr auf »Evidenzbasiertheit« beruft, tarnt diese Überzeugung als neutrales Bekenntnis zu einer besseren, im Trend liegenden Methodologie, und das fällt das den meisten Betrachtern nicht auf.

Wie mit dem Ausdruck »Evidenzbasiertheit« als Qualitätssiegel umgegangen wird, suggeriert bei den Empfängern der Botschaft, dass die Schlussfolgerungen aus präsentierten Daten über jeden Zweifel erhaben seien, ohne dass das dabei explizit so gesagt wird. Wer für seine Arbeit das Attribut »Evidenzbasiertheit« in Anspruch nimmt, kann sich daher bei Kritik jederzeit auf die vergleichsweise bescheidene sachliche Definition von Sackett et al. zurückziehen und dem Kritiker Wind aus den Segeln nehmen. Das semantische Doppelspiel mit den beiden Bedeutungen »über jeden Zweifel erhaben« vs. »das Beste, das beim derzeitigen Wissensstand möglich ist«, erlaubt es dem eloquenten Forscher, Ergebnisse mit weit höherem Gültigkeitsanspruch anzupreisen, als diesen zukommt, ohne Gefahr zu laufen, massiv kritisiert zu werden (Uhl 2007).

Welche Maßnahmen die Gesellschaft bzw. Einzelne mit präventiver Zielsetzung setzen dürfen, sollen bzw. müssen, ist eine ethische Entscheidung, die sich primär aus dem zugrundeliegenden Welt-, Menschen- und Gesellschaftsbild ableitet. Der Begriff »evidenzbasierte Politik« suggeriert allerdings, dass Maßnahmenentscheidungen aus empirischen Fakten abzuleiten seien, und verschleiert so den zentralen Stellenwert der Ethik bei vielen praxisrelevanten Grundsatzentscheidungen. Die vermutete oder nachgewiesene Effektivität einer Maßnahme ist zwar notwendige, aber keinesfalls hinreichende Bedingung, um diese in Erwägung zu ziehen. Die Idee, dass man aus Fakten ableiten könne, was zu geschehen hat, wird bezugnehmend auf Hume (1740) und Moore (1903) als »Sein-Sollen-Fehlschluss« bzw. »Naturalistischer Fehlschluss« bezeichnet.

2.2 Partizipativ-emanzipatorischer vs. paternalistisch-kontrollierender Zugang

Eine zentrale ethische Frage in der Suchtprävention ist, ob der Fokus, im Sinne eines partizipativ-emanzipatorischen Zugangs, primär auf Information, Partizipation und Emanzipation gelegt wird, oder ob, im Sinne eines paternalistisch-kontrollierenden Zugangs, primär auf Kontrolle und Zwang gesetzt wird.

Der partizipativ-emanzipatorische Zugang stimmt mit dem Gesundheitsförderungsansatz im Sinne der Ottawa Charta (WHO 1986a) und dem Gesundheitsbegriff der WHO (1946) überein. Das diesem Ansatz entsprechende Menschenbild baut im Sinne moderner sozialpsychologischer Theorien auf der Überzeugung auf, dass die meisten Menschen in der Lage sind, für sich und ihr Umfeld autonom richtige Entscheidungen zu treffen, wenn man sie darin unterstützt, Lebenskompetenz zu entwickeln, wenn man sie umfassend und ausgewogen informiert, wenn man sie ermutigt, Entscheidungen zu treffen und wenn man sie anleitet, ein zufriedenes Leben anzustreben sowie mit Risiken sinnvoll umzugehen.

Die diesem Zugang diametral entgegenstehende Tendenz, Menschen nicht mittels Empowerment und Information, sondern durch mehr oder weniger massiven Druck dazu zu bringen, bestimmte normative Vorgaben darüber, wie man das Leben gestalten sollte, anzunehmen, widerspricht dem Ansatz der Ottawa Charta und ist mit den Grundsätzen einer demokratischen Gesellschaftsordnung nur schwer vereinbar. Zu dieser Unvereinbarkeit gehören auch Bestrebungen, Gesundheit zum vorrangigen Lebensinhalt hoch zu stilisieren (Healthism, vgl. WHO 1986b).

Um nicht missverstanden zu werden: Selbstverständlich kann es bezüglich des Zuganges in Erziehung, Prävention und Politik kein radikales »entweder – oder« geben. Wenn das Leben oder die Gesundheit der handelnden Personen oder Dritter massiv gefährdet sind und alle anderen Zugänge gescheitert sind, kommt man nicht völlig ohne kontrollierende und sanktionierende Strategien als ultima ratio aus, aber der Schwerpunkt sollte klar auf partizipativ-emanzipatorischen Strategien liegen.

In den letzten Dekaden hat es in der westlichen Pädagogik und Suchtprävention ein klares

Bekenntnis zu partizipativ-emanzipatorischen Strategien gegeben. In jüngster Zeit sind jedoch unübersehbare Anzeichen in Richtung Restauration eines Paternalismus zu erkennen. Die aus den USA kommende starke Anti-Nikotinbewegung, die sich zunächst primär über den Nichtraucherschutz legitimierte, wird zusehends zu einer immer radikaleren Bewegung, die Raucher nicht bloß zu überzeugen sucht, sondern durch unterschiedliche Maßnahmen zur Aufgabe ihres Verhaltens drängen möchte. Auch im Alkoholbereich nimmt die Tendenz zu, nicht bloß auf den Problemkonsum zu fokussieren, sondern jeglichen Alkoholkonsum zu bekämpfen. Einige von der traditionell restriktiven nordeuropäischen Alkoholpolitik geprägte Publikationen, ganz besonders die Bücher »Alkohol – Kein gewöhnliches Konsumgut« von Babor et al. (2005) und "Alcohol in Europe: A Public Health Perspective" von Anderson & Baumberg (2006) haben die Vorstellung popularisiert, dass nur massive Einschränkungen der Alkoholverfügbarkeit über Preiserhöhungen, kürzere Öffnungszeiten, eine Beschränkung der Anzahl der Geschäfte, die Alkohol verkaufen, und der Lokale, die Alkohol ausschenken, kostengünstig und wirksam seien, während Suchtprävention und Suchtbehandlung als teuer und wenig wirksam diskreditiert werden. Die Zukunft wird weisen, ob sich dieser Trend wieder abschwächt oder ob es zu einer Restauration des bereits überwunden geglaubten Paternalismus in Erziehung, Prävention und Politik kommt.

2.3 Der Einfluss der Forschungsfinanziers auf die Ergebnisse

Ein ethischer Grundsatz der Wissenschaft ist auch, dass Forscher der Erkenntnis verpflichtet sein sollten und sich weigern sollten, Ergebnisse zu manipulieren oder einseitig darstellen, um Partikularinteressen von Geldgebern zu bedienen. Wer Grundlagenforschung oder angewandte Forschung – und Evaluation ist angewandte Forschung – finanziert, hat naturgemäß einen mehr oder weniger großen Einfluss auf die Ergebnisse. Das passiert nicht nur, indem mehr oder weniger explizite Vorgaben darüber gemacht werden, was herauskommen soll, sondern auch, indem gezielt solche Forscher ausgewählt und finanziert werden, deren Überzeugungen und Resultate mit den Interessen der Auftraggeber übereinstimmen. Die Sorge, dass die Verlässlichkeit von Forschung durch externe Einflüsse beeinflusst wird, die direkt oder indirekt mit der Finanzierung zusammen hängen, ist ohne Frage begründet. Bestrebungen, den Einfluss der Geldgeber auf Forschungsergebnisse zu begrenzen bzw. transparent zu machen, sind zur Erzielung möglichst verlässlicher Ergebnisse wünschenswert. In diesem Sinne ist auch das "Farmington Consensus Statement" (Davis 1997) zu begrüßen, das von den Herausgebern von 20 Suchtzeitschriften formuliert wurde. Diese schreibt den Autoren Angaben über mögliche Interessenskonflikte ("Conflict of Interest Statements") vor. Auch Bühringer & Batra (2004) sprechen sich gegen die kategorische Ablehnung von Industriesponsering aus und formulieren Rahmenbedingungen, wie die Unabhängigkeit der Forschung im Falle von Geldern aus der Wirtschaft gesichert werden kann.

Die Art und Weise, wie derzeit mit diesen Interessenskonflikten umgegangen wird, ist allerdings hochgradig einseitig und irreführend. Obwohl sich Interessenskonflikte aus finanziellen, persönlichen, politischen und akademischen Aspekten ergeben können (Babor & McGovern 2008), wird der Fokus bei "Conflict of Interest Statements" bei Vorträgen und Publikationen fast ausschließlich auf Wirtschaftsfinanzierung gelegt und so indirekt verschleiert, dass auch andere Finanzgeber und persönliche Motive der Forscher Art und Qualität der Ergebnisse maßgeblich determinieren. Auch Regierungen und NGOs sind oft alles andere als neutral, wenn es darum geht, Forschung zu finanzieren. So war es lange Zeit in den USA unzulässig, Projekte öffentlich zu fördern, wenn die Worte »Abtreibung« oder »Harm Reduction« vorkamen; Antitabak-Initiativen oder Alkoholabstinenzgruppen sind ebenfalls nur schwer davon zu überzeugen, Forscher zu fördern, die ihren Ideen kritisch gegenüberstehen und unter Umständen zu Ergebnissen gelangen, die ihre Überzeugungen konterkarieren. Ein Forscher, der eine wissenschaftliche Karriere machen möchte und nach der Anzahl der Artikel in Peer-Review-Journalen oder nach Impactpunkten beurteilt wird, kann es sich nicht leisten, schlecht abzuschneiden, was opportunistische Anbiederung an Journalredaktionen, Oberflächlichkeit und mitunter sogar Betrug begünstigt. Den meisten Menschen fällt es im Falle solcher intra-

personellen Ambivalenzen aus dissonanztheoretischen Gründen schwer, mit Fakten, die gegen ihre eigenen politischen und ethischen Überzeugungen sprechen, adäquat umzugehen. Die bekannte Poppersche (1934) Maxime, sich ständig selbst kritisch zu hinterfragen, wird in der Forschung bei weitem nicht allgemein praktiziert.

Es ist sicherlich wichtig, Transparenz über die Geldgeber – die Betonung liegt hier auf **alle** Geldgeber – und deren Interessenlagen zu fordern und bei der Beurteilung auch Grundhaltungen der Autoren zu kennen und gegebenenfalls zu berücksichtigen, aber primär geht es um die Qualität des Forschungszuganges, die Nachvollziehbarkeit der Auswertung und die Schlüssigkeit der Interpretationen. Forschung kann nur stattfinden, wenn die notwendigen Mittel dafür vorhanden sind. Da Basisfinanzierung zusehends abnimmt und als Alternative allerorts Drittmittelforschung gefordert wird, ist die einseitige Feindseligkeit gegen wirtschaftlich geförderte Forschung ungerechtfertigt bzw. kontraproduktiv. Rothman (1993) kritisiert den einseitigen und intensiven Kampf gegen von der Wirtschaft finanzierte Forschungsergebnisse zu Recht als "New McCarthyism in Science".

3 Die Grenzen des Wirksamkeitsnachweises

Wir können nicht umhin, Entscheidungen zu treffen und diese fallen besser aus, wenn wir die Auswirkungen unseres Handelns möglichst präzise antizipieren können; also wenn wir auf gut fundierte Ursache-Wirkungs-Hypothesen zurückgreifen können. Aber selbst wenn unsere Prognosen bloß auf schwachen Beinen stehen, können wir die meisten Entscheidungen weder bis zur Klärung aufschieben noch gänzlich unterlassen. Hayek (1988) formulierte treffend: »*Wenn wir alle Handlungen unterließen, für die wir den Grund nicht kennen oder die wir nicht rechtfertigen können, wären wir wahrscheinlich bald tot.*« Realistischerweise müsste man das Wort »wahrscheinlich« in diesem Satz durch »sicherlich« ersetzen.

3.1 Kausalität und nichtexperimentelles Forschen

Erklärtes Ziel der empirischen Wissenschaften ist es, Ursache-Wirkungshypothesen zu formulieren und deren Gültigkeit dann zufallskritisch zu prüfen. Dabei steht der empirische Forscher vor einem grundlegenden Dilemma:

(1) Um Prognosen darüber formulieren zu können, wie sich Maßnahmen auswirken werden, braucht man Kausalmodelle.
(2) Die logisch einwandfreie Überprüfung von Kausalhypothesen ist nur experimentell möglich.
(3) In den Naturwissenschaften passiert dies in der Regel, indem man randomisierte, kontrollierte Studien (RCTs) durchführt, was in den Sozial- und Humanwissenschaften jedoch meist an unüberwindbaren ethischen, ökonomischen, technischen und ontologischen Erkenntnisgrenzen scheitert.

Im Alltag wird uns dieses Dilemma meist nicht bewusst, was Gigerenzer (2002) als zentrale Eigenschaft des Menschen interpretiert: »*Gewissheit zu erlangen, ist offenbar ein grundlegendes Bestreben des menschlichen Geistes. Unsere visuelle Wahrnehmung spiegelt diese Tendenz wider. Ohne dass wir uns dessen bewusst sind, erzeugt unsere Wahrnehmung aus Ungewissheit automatisch Gewissheit.*«

Der sogenannte »Cum-hoc-Fehlschluss«, also der logisch nicht gerechtfertigte Schluss von empirischen Zusammenhängen auf Kausalität (Pirie 1985, Curtis 2006), ist demnach nicht Ergebnis unlogischen Denkens, sondern ein Täuschung im Sinne eines Gestaltphänomens, die spontan in unseren Köpfen entsteht. Wie Kritz et al. (1980) formulierten: »*Die geforderte ›reine Beobachtung‹ ergibt sich also paradoxerweise erst als Zerlegung eines unmittelbaren Wahrnehmungserlebnisses durch eine gedankliche (analytische) Leistung, also letztlich eine ›Interpretation‹ aufgrund von Wissen.*«

Die Welt mit den Methoden der Wissenschaft zu verstehen, bedeutet demnach nicht, »reine Beobachtungen« korrekt zu interpretieren, sondern die ohne bewusstes Zutun entstandenen Zusammenhangsvorstellungen, unter Bezugnahme auf vorhandenes Wissen, zunächst zu dekonstruieren und dann – im Sinne von Popper (1934), darauf aufbauend – kritisch nach alternativen Erklärungen

zu suchen sowie sich anschließend systematisch auf die Suche nach Pro- und Kontraargumenten für die jeweiligen Erklärungsansätze zu machen.

Will man aus nicht experimentell gewonnenen empirischen Zusammenhängen logisch korrekt Kausalhypothesen ableiten, so bedarf es mehr oder weniger plausibler Zusatzannahmen. Meist gibt es mehrere plausible bzw. zumindest denkbare Optionen, auf die aufbauend sich bedingte hypothetische Schlussfolgerungen über Ursache-Wirkungs-Zusammenhänge ableiten lassen (»Wenn das und das zutrifft, dann …«). Nur wenn diese Annahmen explizit getroffen und dokumentiert werden, sind präzise Diskurse über den Gegenstand und eine Weiterentwicklung des Erkenntnisstandes möglich. Die für den empirischen Forscher äußerst unerfreuliche Konsequenz, dass empirische Forschung nicht primär auf empirische Fakten aufbauen kann, sondern immer auch Annahmen getroffen werden müssen, hat Feyerabend (1978) auf den Punkt gebracht: »*Wir erkennen langsam – und ohne die Situation völlig zu verstehen – dass alle Argumente zugunsten realistischer Alltagsüberzeugungen und wissenschaftlicher Theorien sich im Kreis drehen; Sie nehmen an, was sie beweisen wollen.*«

Da die meisten Rezipienten von eigentlich deskriptiven Darstellungen diesen spontan und ohne sich der Problematik bewusst zu werden, Kausalität, und damit eine Zusammenhangsbedeutung, einhauchen, werden empirische Befunde, die zunächst wenig zur Klärung relevanter Fragestellungen beitragen können, nur selten offen kritisiert. Auf diesen Mechanismus schielend formulierte Huff (1954) die zynische Maxime: »*Wenn du nicht beweisen kannst, was du beweisen möchtest, so beweise etwas anderes und benimm dich so, als wäre es das Gleiche.*[3]«

Anstatt starr die Trampelpfade des State-of-the-Art zu beschreiten, indem man die Forschungsstrategien anerkannter Forscher mechanistisch kopiert, oder sich im Wissen, dass die Leser Ergebnisse überinterpretieren werden, auf reine Beschreibung zu beschränken, sollte man sich den grundlegenden Forschungsproblemen stellen, implizite Annahmen explizieren und zur Kenntnis nehmen, dass vieles mehrdeutig, unsicher und spekulativ ist – und wohl auch bleiben wird. Es geht eben darum, aus einer undurchschaubaren, komplexen Situation das Bestmögliche herauszuholen. Hartnoll (2004) meinte dazu: »*Nach dem impliziten Verständnis von Wissenschaft handelt es sich um einen Prozess, in dem sich relevante Fragestellungen ergeben, wo vorhandene Fakten, wie bei einem Puzzle, zusammengesetzt werden, wo fehlende Teile, basierend auf Common-Sense und Logik, temporär ergänzt werden und wo gegebenenfalls gezielt durch weitere Forschungsschritte nach Einsichten gesucht wird. In diesem Sinne ähnelt der Wissenschaftler einem Detektiv, der systematisch Indizien sammelt und zusammenstellt, bis der Fall gelöst ist.*« Diederich (1974) meinte dazu: »*Die Einsicht, dass fast alle wissenschaftlichen Theorien falsch sind, führt zur Notwendigkeit, wissenschaftlichen Fortschritt mit Hilfe einer Charakterisierung von Theorien zu begreifen, die zwischen besseren und schlechteren, wenngleich allesamt falschen Theorien, zu entscheiden gestattet*«.

In letzterem Sinn werden nun, ohne Anspruch auf Vollständigkeit, einige im Zusammenhang mit der Evaluation von Prävention relevante Problemfelder angerissen und abschließend darauf aufbauend Schlussfolgerungen für eine sinnvolle Evaluationspraxis gezogen.

3.2 Komplexität

Wir alle sind ständig unterschiedlichen Einflüssen ausgesetzt, die wir großteils gar nicht bewusst wahrnehmen, wenn wir sie wahrnehmen, kaum adäquat messen können, und als Forscher in der Regel auch nicht kontrollieren (konstant halten) können. Die einzige realistische Möglichkeit, diese Einflüsse zu kontrollieren, sind sauber geplante, randomisierte Experimente (RCT) mit großen Experimental- und Kontrollgruppen, wo sich die Einflüsse ausmitteln lassen. Wie erwähnt sind solche Studien häufig aus unterschiedlichen Gründen nicht realisierbar. Aber selbst wenn ein ideales Experiment durchführbar ist, ist der direkte Vergleich zwischen Experimental- und Kontrollgruppe meist wenig aufschlussreich, weil sich die relevanten Prozesse in komplexen Zusammenhängen mit vielen beteiligten und nicht berücksichtigten Variablen manifestieren.

Menschen sind durchwegs in der Lage, äußerst komplexe Geschehnisse rasch und gleichzeitig

3 Frei übersetzt von Bosbach & Korff (2011, S. 176). Englischer Originaltext: "The general method is to pick two things that sound the same but are not. (…) you haven't proved a thing, but it rather sounds as if you have, doesn't it?"

zu erfassen – Dörner (1996) spricht in diesem Zusammenhang von »Superzeichen« im Sinne von Gestaltphänomen – und darauf bezugnehmend intuitiv adäquat zu handeln (Gigerenzer 2008). Man denke hier z. B. an die blitzschnell zu treffenden Entscheidungen über das weitere Verhalten inkl. der adäquaten technischen Bedienung, wenn man sich mit einem PKW bei dichtem Verkehr auf einer davor nie gefahrenen Straße einer ungeregelten Kreuzung nähert. Meist, aber nicht immer gelingt die Situationserfassung, in geschildertem Fall an der Unfallhäufigkeit erkennbar. Diese grundlegende Fähigkeit ist keinesfalls geringzuschätzen. Menschen sind ständig mit komplexen Situationen konfrontiert und schaffen es zu überleben. Wenn man als Forscher hingegen versucht, die relevanten Variablen einzeln zu messen und mit quantitativen statistischen Datenanalysemodellen zu verstehen, ist eine vergleichbare Leistung mit an Sicherheit grenzender Wahrscheinlichkeit auszuschließen. Resultate, wie Betagewichte bei multivarianten Analysen oder Faktorenladungen bei Faktorenanalysen werden in der Regel frei assoziierend ad hoc interpretiert. Von wenigen Ausnahmen abgesehen hat das mit fundiertem Forschungszugang wenig zu tun. Die Beliebigkeit derartiger Assoziationen fällt allerdings nur wenigen Betrachtern auf, was mit der erwähnten Bestrebung bzw. Fähigkeit unserer Wahrnehmung, aus Ungewissheit automatisch Gewissheit zu erzeugen, erklärt werden kann, oder aber auch mit einer generell unkritischen Haltung jeglichen mehr oder weniger glaubhaft präsentierten Daten gegenüber.

3.3 Nichtlineare Systeme

Eine implizite Annahme beim experimentellen Zugang ist, dass gesetzte Interventionen weitgehend homogen wirken, also dass jedes Individuum, von gewissen Zufallsschwankungen abgesehen, gleich reagiert. Dieses Modell ist für die meisten pharmakologischen Studien adäquat. Die Einnahme einer bestimmten Tablette bewirkt bei der Mehrzahl der Personen ähnliche Effekte.

Ganz anders verhält es sich aber bei sozialen Interventionen. Kleine Interventionen, die auf die meisten tangierten Personen gar keinen Einfluss haben, können unter bestimmten Rahmenbedingungen enorme Effekte bewirken. Das passiert z. B., wenn im Zuge eines Präventionsprojektes an einer Schule ein einzelner Lehrer von einem Gedanken so sehr angetan ist, dass er beim Direktor aktiv wird und längerfristig eine Veränderung zentraler Strukturelemente der Schule auslöst. Laut Chaostheorie (z. B. Steward 1989) sind unvorhersehbare Ursache-Wirkungs-Zusammenhänge in der Natur weitverbreitet. Pointiert dargestellt wird dieser Sachverhalt durch die Behauptung, dass die Flügelbewegung eines Schmetterlings in Südamerika gegebenenfalls durch eine lange Kette von Abfolgen einen Wirbelsturm in Indonesien auslösen kann (Schmetterlingseffekt). Selbst wenn uns der Schmetterlingseffekt in der ursprünglichen Version eher unplausibel erscheint, haben wir wahrscheinlich kein Problem, uns vorzustellen, dass der gleiche Schmetterling vor dem Fenster eines Atomkraftwerks, wenn er einen Techniker in einem ungünstigen Augenblick ablenkt, eine Megakatastrophe auslösen kann, obwohl Millionen Schmetterlinge, die beim Atomkraftwerk vorbeifliegen, die Geschehnisse dort kein bisschen tangieren.

Dieses Phänomen, das man mit »Generativität« umschreiben kann (Uhl 1998), verursacht bei der Evaluation von Maßnahmen große Probleme. Mitunter überragen derartige unvorhersehbare und nur selten auftretende Phänomene das Ausmaß des durchschnittlichen Effekts bei weitem. Generativität sollte aber nicht mit ursprünglich nicht erwarteten, jedoch – ab dem Zeitpunkt ihres Auftretens – vorhersehbaren systemischen Effekten verwechselt werden, z. B. wenn bestimmte Diskurse in einer Art und Weise stimuliert werden, dass sich die öffentliche Meinung ändert und in der Folge strukturelle Veränderungen initiiert werden. Die Erfassung zunächst nicht erwarteter aber nichtsdestoweniger vorhersehbarer Effekte wird unter den Ausdruck “Impact Evaluation” subsumiert.

3.4 Messprobleme

3.4.1 Programme vs. Fähigkeiten

Manuale, die Präventionsprogramme beschreiben, können für Präventionsfachkräfte recht nützlich sein, aber was in der Praxis wirklich zählt,

ist, was diese Personen aus diesem Rohmaterial machen und wie sie mit der Zielgruppe interagieren. Ob, wie oft und mit wie vielen Teilnehmenden ein bestimmtes Programm angewandt wird, kann leicht festgehalten werden, aber die wichtigen persönlichen Fähigkeiten, die den Erfolg letztlich ausmachen, sind schwer zu erfassen und werden in der Regel völlig ignoriert. Wenn man eher unbedeutende Faktoren (z. B. Tageszeit der Maßnahmendurchführung) kontrolliert und jene Faktoren, die erhebliche Bedeutung haben, nicht (z. B. Emphatievermögen des Durchführenden), ist es wenig wahrscheinlich, dass wichtige Veränderungen korrekt erfasst und auf die relevanten Faktoren attribuiert werden. In einem gewissen Sinn erinnert die Strategie an den Betrunkenen, der die verlorenen Schlüssel unter einer Laterne sucht, weil es dort hell ist, und nicht dort sucht, wo er sie verloren hat.

3.4.2 Messung von Ersatzvariablen statt Zielkriterium

Suchtprävention zielt – in Hinblick auf Substanzen – darauf ab, die Wahrscheinlichkeit für das Auftreten problematischen bzw. süchtigen Substanzkonsums in der Zukunft zu verringern. Relevant ist hier das Zielkriterium »problematischer Substanzkonsum« (der süchtigen Konsum miteinschließt). Da »problematischer Substanzkonsum« in der Mehrzahl der Fälle erst in der weiteren Zukunft auftritt, ist man bei Studien zur Messung von Präventionseffekten auf Ersatzvariablen (Mediatorvariablen, Surrogate Endpoints, Proxy Variables) angewiesen, wenn man den Beobachtungszeitraum nicht auf viele Jahre ausdehnen möchte oder kann. Hierfür geeignet sind Variablen, von denen man begründeterweise annehmen kann, dass diese mit dem eigentlichen Zielkriterium kausal zusammenhängen. »Begründeterweise« bedeutet hier »experimentell« oder »zumindest quasiexperimentell« überprüft und auch das nur, wenn die Rahmenbedingungen beim Experiment mit der Anwendungssituation übereinstimmen. Wird bloß von einem korrelativen Zusammenhang auf Kausalität geschlossen oder Kausalität bloß intuitiv angenommen, so ist Skepsis angezeigt.

Eine populäre Ersatzvariable, um Problemkonsum in der Zukunft indirekt zu erfassen, ist aktueller »unproblematischer Konsum« (von Probierkonsum bis moderater Konsum). Betrachtet man die Entwicklung von Abstinenz zu Problemkonsum in den drei Stadien »Abstinenz«, »unproblematischer Konsum« und »Problemkonsum«, kann logischerweise das dritte Stadium nur erreicht werden, wenn zuvor der Übergang von Stadium 1 zu Stadium 2 erfolgte. Wird jeglicher Konsum kategorisch verhindert, kann logischerweise auch kein Problemkonsum auftreten. Wird durch eine Maßnahme eine deutliche Reduktion des Einstiegs in den unproblematischen Konsums bewirkt, so bedeutete das aber nicht zwangsläufig eine Verringerung des zukünftigen Problemkonsums, wie man leicht zeigen kann.

Aus der Alkoholepidemiologie ist bekannt, dass Menschen mit psychischen, körperlichen und sozialen Problemen verstärkt entweder zu Alkoholabstinenz oder zu problematischen Konsumformen neigen, was sich in einem u-förmigen bzw. j-förmigen Zusammenhang zwischen Problemen und Alkoholkonsum manifestiert (Uhl et al. 2009c). Gelingt es nun durch geeignete Präventionsansätze, Gesundheitsförderungsprogramme, soziale Maßnahmen oder therapeutische Angebote die Problemlast der Bevölkerung deutlich zu verringern, so ist zu erwarten, dass sowohl die Neigung zur völligen Abstinenz als auch die Neigung zum problematischen Alkoholkonsum abnimmt. Weniger Abstinente bedeutet mehr Konsumenten. Dass mehr Konsumenten mit weniger Problemkonsumenten korrespondieren können, mutet zunächst paradox an, ist allerdings nicht paradox, wie eine Analogie zu Verkehrsunfällen unmittelbar zeigt.

Ohne Kraftfahrzeuge kann es keine KFZ-Unfälle mit Personenschaden geben. Die Anzahl der KFZ-Zulassungen ist über die letzten Jahrzehnte um ein Vielfaches gestiegen und trotzdem ist die Anzahl der KFZ-Unfälle mit Personenschaden um 95 % zurückgegangen (Uhl et al. 2009c).

Hierzu die theoretische Erklärung mittels »Übergangswahrscheinlichkeiten«: Maßnahmen, die auf die Übergangswahrscheinlichkeit $p_{1,2}$ (von Stadium 1 zu 2) abzielen, können sich auf die Übergangswahrscheinlichkeit $p_{2,3}$ (von Stadium 2 zu 3) sehr unterschiedlich auswirken. Bei der Beurteilung, wie sich eine Veränderung auf einen bestimmten Parameter (hier Übergangswahrscheinlichkeit $p_{1,2}$) auswirkt, unterliegen wir üblicherweise dem Trug-

schluss, dass wir implizit davon ausgehen, dass die anderen Parameter (hier die Übergangswahrscheinlichkeit $p_{2,3}$) konstant bleiben.

Um nicht missverstanden zu werden: Ich vertrete in diesem Abschnitt nicht die These, dass die Zunahme des unproblematischen Konsums von psychoaktiven Substanzen zwangsläufig zu einer Abnahme des Problemkonsums führt. Ob die Veränderungen beim unproblematischen Konsum mit einer gleichgerichteten, gar keiner oder umgekehrten Veränderung des Problemkonsums einhergehen, hängt von der Art der Intervention und vom Kontext ab. Hier ist theoretisch alles möglich. Festzuhalten ist nur, dass man aus Veränderungen der Prävalenz des »unproblematischen Konsums« (z. B. Probierkonsum bei Jugendlichen) nicht einfach auf Veränderungen in der Prävalenz des späteren »Problemkonsums« schließen kann.

3.4.3 Papageieffekt und Bumerangeffekt – Ist die Messung von Einstellungen überhaupt möglich?

Eine weitere populäre Variable zur indirekten Erfassung des späteren Problemkonsums sind »Einstellungen zum Substanzkonsum«, wenngleich in der Fachwelt der kausale Zusammenhang zwischen Einstellungen und Verhalten oft angezweifelt wird. Die Frage, die sich hier stellt, ist meiner Auffassung nach weniger, ob Einstellungen und Verhalten korrelieren bzw. in einem Kausalverhältnis zueinander stehen, sondern, ob das, was meist gemessen wird, tatsächlich Einstellungen sind.

Viele Kinder und Jugendliche bis zu einem gewissen – individuell unterschiedlichen – Alter haben keine Ahnung davon, was psychoaktive Substanzen sind, interessieren sich nicht für dieses Thema und haben daher auch keine konkrete Einstellung dazu. Wenn nun bei schulischen Präventionsprojekten Schüler vor und nach dem Projekt, bei dem sie mit mehr oder weniger sachlicher Aufklärung, oft jedoch mit übertriebenen Gefahrenzuschreibungen konfrontiert worden sind, aufgefordert werden, ihre Einstellung zu psychoaktiven Substanzen zu skalieren, reproduzieren sie bei der zweiten Messung, wie in einer Prüfungssituation über rezipierten Lernstoff, jene Inhalte, die sie sich gemerkt haben. Die Zustimmung zur Aussage »Drogen sind extrem gefährlich« unterscheidet sich hier kaum von der Zustimmung zu »Paris ist die Hauptstadt von Frankreich« oder »Sieben ist eine Primzahl«. Auch Schülern, die bereits Erfahrungen mit psychoaktiven Substanzen gemacht haben und die nicht vorhaben, ihre Experimente nunmehr zu unterlassen, wird hier rasch klar, in welchem Sinne ihre Antwort erwünscht ist und verhalten sich beim Fragebogenausfüllen dementsprechend.

Zu erwarten, dass jemand, der ohne Bezug zur Thematik oder mit einem schulischen Kontext im Hintergrund den Satz »Drogen sind extrem gefährlich« reproduziert, in Zukunft keine Drogen konsumieren wird, ist ähnlich naiv, wie die Überzeugung, dass ein Papagei, dem man beibringt zu sagen »Ich hasse Bananen!«, eine angebotene Banane ablehnen wird . Wenn Kinder ohne Bezug zum Thema »überzogene Gefahrenaussagen« reproduzieren (Papageieffekt, Uhl 2002) sollte das nicht als Einstellung interpretiert werden. Echte Einstellungen entstehen in der Regel erst, wenn die Betreffenden, meist im Freundeskreis, mit der Materie real konfrontiert werden. Wenn sie dann allerdings merken, dass das, was sie beobachten und erleben, mit der erfolgten »Aufklärung« nicht übereinstimmt, verlieren sie den Glauben an erwachsene Aufklärer und verlassen sich primär auf Peergroup-Informationen, die die tatsächlichen Gefahren oft unterschätzen.

Erwachsene erleben es in der Regel als Erfolg, wenn ihre überzeichneten Gefahrenurteile von Zielpersonen mechanisch reproduziert werden, weil sie die Aussagen als Einstellungen interpretieren und nicht ahnen, dass sie eigentlich ihre Glaubwürdigkeit untergraben und längerfristig das Gegenteil von dem bewirken, was ihr Ziel war (Bumerangeffekt). Der Papageieffekt verschleiert kurzfristig, dass langfristig tatsächlich ein Bumerangeffekt induziert wurde.

3.4.4 Die Vagheit verbaler Ausdrücke

Sprache ist unvermeidbar vage und mehrdeutig. An einer Kommunikation beteiligte Personen müssen die Bedeutung der Mitteilungen ihres Gegenübers unter Bezugnahme auf den Kontext interpretieren, mehr oder weniger schwere Missverständnisse sind dabei unvermeidlich. Sehr tref-

fend formulierte das Luhmann (2000), der meinte: *»Als erstes ist unwahrscheinlich, dass einer überhaupt versteht, was der andere meint, gegeben die Trennung und Individualisierung ihres Bewusstseins. Sinn kann nur kontextgebunden verstanden werden, und als Kontext fungiert für jeden zunächst einmal das, was sein eigenes Gedächtnis bereitstellt.«* Uns sind die sprachlich bedingten Missverständnisse in der Regel überhaupt nicht bewusst und wir glauben, von Ausnahmen abgesehen, die anderen zu verstehen, uns eindeutig auszudrücken und verstanden zu werden. Das übliche Gefühl, dass das, was man gesagt bzw. gehört hat, eigentlich nur eine bestimmte Bedeutung haben kann, drückte Liev Schreiber in seinem Film "Everything Is Illuminated" (2005) sehr treffend aus: Als der Reiseführer Alex vom Hauptdarsteller gefragt wird, was er mit einer Aussage meine, antwortet dieser: »Genau was ich gesagt habe! Hätte ich etwas anderes ausdrücken wollen, hätte ich etwas anderes gesagt!«.

Das Gefühl, uns meist klar auszudrücken und präzise zu verstehen, was andere meinen, macht das Leben zwar subjektiv einfacher, stellt sich aber in der Wissenschaft oft als Problem dar, weil die Klärung von semantischen Schwierigkeiten deutlich erschwert wird, wenn alle davon überzeugt sind, dass es gar kein Verständnisproblem gibt. Besonders problematisch ist das in Zusammenhang mit Fragebogenerhebungen, wo die Entwickler von ihrer Perspektive aus glauben, eindeutig zu formulieren, jeder, der den Fragebogen ausfüllt, die Fragen ad hoc ganz anders interpretiert, aber das Gefühl hat, adäquat auf die Fragestellung zu reagieren, und jene, die die Ergebnisse auswerten, ebenfalls davon überzeugt sind, die wahre Bedeutung der Fragen zu erfassen. Kaum ein Proband zögert, z. B. die Aussage »Ein gewisses Maß an Aggressivität ist positiv!« in Sekundenbruchteilen auf einer Skala von »trifft zu« bis »trifft nicht zu« zu beurteilen, ohne sich zu fragen, wie viel ein »gewisses Maß« ist, was mit »Aggressivität« hier gemeint ist bzw. für wen es positiv sein soll.

Wie sich das in der Praxis auswirkt, zeigt recht ernüchternd eine im Anschluss an die österreichische ESPAD-Erhebung 2008 durchgeführte qualitative Validierungsstudie an 100 zufällig ausgewählten 14- bis 17-jährigen Schülern, die zuvor den Fragebogen ausgefüllt hatten. Diese wurden darüber befragt, wie sie einzelne Fragen interpretiert und beantwortet hatten. Dabei stellte sich in Bezug auf eine einfache Frage zum letzten Alkoholkonsum heraus, dass sie von 16 Personen völlig missverstanden worden war, dass vier die Frage wahrscheinlich falsch verstanden hatten, drei sich nicht die Mühe gemacht hatten, sie verstehen zu wollen, und irgendetwas geantwortet hatten, 38 die Frage zwar richtig verstanden hatten, aber die korrekte Antwort nicht wussten, und dass nur knapp über einem Drittel (39 Schüler) die Frage richtig verstanden und ihrer eigenen Meinung nach korrekt beantwortet hatten (Schmutterer et al. 2008). Bei anderen untersuchten Fragen ergaben sich ähnlich haarsträubende Validierungsergebnisse. Zur Untermauerung, dass man sich über derartige Ergebnisse eigentlich nicht wundern braucht, kann man die PISA-Studie heranziehen, die ergeben hat, dass in dieser Altersgruppe gut 20 % große Schwierigkeiten mit sinnverstehendem Lesen haben (Reiter 2005) und dass in Schulsettings derartige Erhebungen vielfach überhaupt nicht ernst genommen werden. Was verwundert, ist allerdings, dass ESPAD und vergleichbare Erhebungen in der Fachwelt nach wie vor sehr ernst genommen und auf Prozentpunkte genau interpretiert werden.

3.4.5 Die Unvermeidbarkeit impliziter Kommunikation

Zum eben genannten Problem mit der gegebenen Vagheit und Mehrdeutigkeit der Sprache, insbesondere in Zusammenhang mit Fragebogenerhebungen und interviewergestützten Befragungen, kommt noch das Problem, dass man die Ergebnisse mit etwas Geschick – oder auch mit lauterer Absicht völlig unbewusst – durch die implizite Kommunikation über die Wortwahl umfassend beeinflussen kann. Ohne Validierungsstudien, die in diesem Bereich jedoch kaum üblich sind, kommt man der Frage nach einer realitätsgetreuen Abbildung von zu erfassenden Phänomenen nicht wirklich nahe. Wenn man die Erfahrungen mit einer psychoaktiven Substanz abfragt und dabei die Kategorien »nie«, »ein- oder zweimal«, »drei- bis fünfmal«, »sechs- bis zehnmal« sowie »elfmal oder mehr« anbietet, teilt der Verfasser des Fragebogens implizit mit, dass er »mehr als elfmal« bereits als sehr viel erachtet. Wer hingegen die Kategorien »nie«, »ein- bis zwanzigmal«, »einundzwanzig- bis hundertmal« sowie »hundertmal und mehr«

anbietet, der zeigt, dass er elfmal als relativ wenig erachtet. Da sich viele Befragte im Sinne der sozialen Erwünschtheit an solchen Vorgaben orientieren, wirken sich diese stark auf das Ergebnis aus.

Ein besonders deutliches Beispiel für implizite Kommunikation konnte ich in meiner Anfangszeit als Forscher beobachten. Ich fragte einen Mann, wie er zur Entkriminalisierung des Haschischkonsums stehe, und statt adäquat zu antworten, sagte er bloß: »Sehr interessant«. Auf Nachfrage wiederholte er neuerlich: »Interessant«. Danach führte er aus: »Wenn sie als Wissenschaftler in einem seriösen Forschungsinstitut mich so etwas fragen, bedeutet das, dass Haschisch nicht so gefährlich sein kann, wie ich bis jetzt vermutete. Dann hat mein Sohn also Recht. Dann kann man es vielleicht wirklich legalisieren.«

In diesem Zusammenhang sollte man sich im Klaren sein, dass es wesentlich leichter ist, das Verbalverhalten gezielt oder unabsichtlich zu manipulieren als das Konsumverhalten selbst. Wenn man als Interviewer durchblicken lässt, dass die vertrauliche Behandlung der Fragebögen unter Umständen nicht gewährleistet ist und die dort gemachten Angaben über illegale Verhaltensweisen negative Konsequenzen nach sich ziehen könnten, ist damit zu rechnen, dass die selbstberichtete Prävalenz deutlich niedriger als die tatsächliche ausfällt.

3.4.6 Unreliable Selbstangaben

In der Regel ist es aus finanziellen oder praktischen Gründen nicht möglich, den Substanzkonsum verlässlich zu erheben, und man verlässt sich auf Selbstangaben. Die Anzahl der Fehlerquellen, die sich hier auswirken können, ist groß. Die Befragten müssen die Frage verstehen, die korrekte Antwort wissen, motiviert sein, die korrekte Antwort zu geben und in der Lage sein, diese verbal oder in den Kategorien eines Fragebogens auszudrücken (Alvin 2007). Dass da viel schiefgehen kann, bedarf keiner näheren Erörterung und dass da viel schief geht, zeigt sich oft, wenn man kritische Vergleiche anstellt. So ergaben zwei österreichische Repräsentativerhebungen im Abstand von vier Jahren einen Rückgang der Lebenszeitprävalenz des Cannabiskonsums, also des Anteils jener, die bereits mindestens einmal Cannabis probiert hatten, von 18 % auf 12 %, was logisch unmöglich ist. Es ist zwar möglich, dass die Zahl jener, die im letzten Monat oder letzten Jahr Cannabis geraucht haben, von einer Erhebung zur nächsten stark zurückgeht, aber jemand, der diese Erfahrung bereits gemacht hat, kann sie nie wieder verlieren. Im Verlauf von 4 Jahren sind, infolge der Bevölkerungsdynamik, Schwankungen um 1 % die Lebenszeitprävalenz betreffend denkbar, wobei eine geringfügige Zunahme weit plausibler ist, als eine geringfügige Abnahme.

Wie dieser artifizielle Rückgang entstanden ist, kann man nur mutmaßen. Unter Umständen hat ein relevanter Anteil der Cannabiserfahrenen infolge von Kampagnen bzw. öffentlichen Diskussionen, die Cannabis gefährlicher darstellten, als das lange Zeit der Fall war, beschlossen, ihre diesbezüglichen Erfahrungen fremden Interviewern gegenüber nunmehr zu verschweigen (Uhl et al. 2009c). Ähnlich widersprüchliche und unplausible Ergebnisse gibt es auch den Zigarettenkonsum in Österreich betreffend (Uhl et al. 2009a).

3.4.7 Eigentlich geht es in der Prävention um Langzeiteffekte

Bei der Evaluation von Präventionsmaßnahmen wird oft gesagt: »Kurz nach der Intervention lassen sich Effekte objektivieren, aber längerfristig zeigen sich keine Effekte.« Das wird als Erfolgsnachweis interpretiert. Professionelle Suchtprävention, Erziehung, Unterricht u. v. m. zielen darauf, die Chancen der Zielpersonen langfristig dahingehend zu verändern, dass diese ihr Leben erfolgreicher und freudvoller gestalten können – das ist eindeutig eine Langzeitperspektive. Dass begrenzte Maßnahmen nur begrenzte Effekte haben können und dass diese im Konzert der unterschiedlichen Einflüsse nicht zufallskritisch belegbar sind, ist argumentierbar. Es zählt letztlich der Gesamteffekt aller Maßnahmen – aber zu argumentieren, dass diese Maßnahmen nach eigener Überzeugung keine langfristigen Effekte haben oder haben können und trotzdem auf deren Wirksamkeit zu bestehen, mutet grotesk an.

3.5 Methodologische Probleme

3.5.1 *Präventionsforschung vs. Evaluation von konkreten Anwendungen*

Präventionsforschung beschäftigt sich mit der Frage, ob bestimmte Präventionszugänge sinnvoll bzw. wirksam sind. Hier geht es um die Evaluation von Techniken, Methoden, Programmen etc. Wenn diese angewandt werden, erwartet man von »Evaluation« hingegen, dass sie die Qualität der praktischen Umsetzung prüft und nicht die Wirksamkeit neuerlich belegt, was in der Regel auch gar nicht geht.

In der Medizin führt der Umstand, dass beide Forschungszugänge als »Evaluation« bezeichnet werden können, kaum zu Verwirrung. Die Evaluation der Wirksamkeit von Medikamenten erfolgt in pharmakologischen Studien. Und kaum jemand, der eine medizinische Ambulanz evaluieren soll, kommt auf die Idee zu untersuchen, ob die dort verabreichten Medikamente tatsächlich wirksam sind. Bei der Evaluation einer Ambulanz geht es darum zu prüfen, ob das Personal motiviert ist, ob Indikationen korrekt gestellt werden, ob Dosen adäquat verschrieben werden, ob Hygienestandards eingehalten werden, ob Patienten korrekt informiert werden u. v. m.

Bei der Evaluation von Suchtpräventionsprojekten ist die Idee, dass man deren Wirksamkeit zufallskritisch nachweisen müsse, hingegen allgegenwärtig. Dieses Unterfangen ist angesichts begrenzter Stichprobenumfänge und der dadurch begründbaren geringen statistischen Macht meist aussichtslos. Wenn dieser Versuch erwartungsgemäß scheitert, ist die Frustration groß.

3.5.2 *Effektgrößen, Inzidenz, "Number Needed to Treat" (NNT) und notwenige Stichprobenumfänge*

Präventive Maßnahmen sind üblicherweise bloß zeitlich begrenzte Einzeltöne in einem Orchester anhaltender und konkurrierender Einflüsse. Wenn nun ein solches Präventionsprogramm die Auftrittswahrscheinlichkeit (Inzidenz) für ein bestimmtes Problem um 20 % reduziert (Relativ Risk Reduction, RRR = 20 %), ist das ein gewaltiger Erfolg. Derzeit können wir davon ausgehen, dass rund 1 % der Bevölkerung im Laufe des Lebens ein ernstes Problem mit illegalen Drogen entwickelt (Gesamtlebenszeitprävalenz). Da die Manifestation irgendwann im Laufe des Lebens passiert, ist bei einer Lebenserwartung von 80 Jahren mit einer jährlichen Inzidenz von 0,01 % zu rechnen und man kann annehmen, dass auch in sehr gefährdeten Altersabschnitten die Inzidenz pro Jahr maximal 0,1 % beträgt.

Man kann also erwarten, dass ohne Präventionsmaßnahmen pro Jahr maximal eine von 1.000 Personen in gefährdeten Altersabschnitten dieses Problem entwickeln wird (Annual Control Event Risk, CER = 0,1 %). Bei einer relativen Risikoreduktion um 20 % bedeutet das, dass in der mit dem Präventionsprogramm konfrontierten Gruppe durchschnittlich 0,8 von 1.000 Personen das Problem entwickeln (Annual Experimental Event Risk, EER = 0,08 %). Die absolute Risikoreduktion (Absolut Risk Reduction, ARR) durch das Programm macht damit 0,2 von 1.000 Fällen aus; oder anders formuliert, das Präventionsprogramm müsste an 5.000 Personen durchgeführt werden (Number Needed to Treat, NNT = 5.000), um innerhalb eines Jahres die Problemmanifestation bei **einer** Person zu verhindern (Doll 2008).

Wenn wir nun von einer "Number Needed to Treat" von 5.000 Personen ausgehen, die statistische Power auf 80 %[4] und den Alphafehler auf 5 % festlegen[5], so benötigen wir jeweils knapp 300.000[6] Personen in der Kontrollgruppe und in der Versuchsgruppe, also eine Gesamtstichprobe von fast 600.000 Personen (Uhl 2010), um ein statistisch signifikantes Ergebnis zu liefern. Da Präventionsprogramme meist in Gruppen abgehalten werden (Klumpenstichproben statt echter Zufallsstichproben) und die Identifikation von Problempersonen nicht völlig fehlerfrei möglich ist, müsste man die notwendige Gesamtstichprobe noch deutlich höher ansetzen.

Es bedarf wohl keiner weiteren Erläuterungen, um klar zu machen, dass Strichprobenumfänge in diesen Dimensionen die Möglichkeiten jeder Forschungseinrichtung erheblich überschreiten.

4 d. h. den Effekt mit mindestens 80 % Wahrscheinlichkeit nachzuweisen möchten.

5 d. h. in Kauf zu nehmen, dass in 5 % der Fälle Zufallssignifikanzen auftreten, selbst wenn es tatsächlich gar keine Effekte gibt.

6 Genau errechnet wären das 287.877 Personen

3.5.3 Statistische Regressionsartefakte – Pseudoveränderungen

Kahneman (2012) berichtet von einem militärischen Ausbildner, der Kadetten, die Höchstleistungen bringen, nicht mehr lobte, aber jene, die schlechte Leistungen brachten, kritisierte. Er hatte nämlich beobachtet, dass jene, die bei einer Gelegenheit besonders gut abschnitten, als er sie lobte, beim nächsten Mal schlechter abschnitten. Er hatte auch beobachtet, dass jene, die bei einer Gelegenheit besonders schlecht abschnitten, als er sie kritisierte, beim nächsten Mal besser abschnitten. Kahneman erklärte dem Ausbilder, dass dieser einem klassischen Regressionsartefakt aufgesessen sei und dass dieser Effekt ähnlich auch aufgetreten wäre, wenn er auf die einzelnen Leistungen gar nicht reagiert hätte. Aus der Sozialpsychologie ist bekannt, dass – ganz im Gegensatz zur intuitiven Schlussfolgerung des Ausbilders – Lob langfristig Leistungen verbessert und Tadel diese verschlechtert. Die Erklärung für den Artefakt ist einfach: Es gibt bei allen Personen Leistungsschwankungen. Wenn langfristig alles völlig unverändert bleibt, sind jene, die einmal besonders gut abschneiden, im Durchschnitt beim nächsten Mal schlechter, und jene, die einmal besonders schlecht abschneiden, im Durchschnitt beim nächsten Mal besser, ganz gleich, wie darauf reagiert wird.

Regressionsartefakte entstehen durch inhaltlich irrelevante Schwankungen über die Zeit (Generalisierungsfehler = verringerte Validität, vgl. dazu »Theorie der Generalisierbarkeit« nach Cronbach et al. 1972) und durch messfehlerbedingte Variabilität (reduzierte Reliabilität) immer dann, wenn eine Extremstichprobe anhand einer Variable selektiert wird, die von der Zielvariable stochastisch abhängig oder mit dieser gar identisch ist. Dieser Artefakt tritt daher auf, wenn man z. B. Kinder, die bei einem Test besonders gut abgeschnitten haben, neuerlich testet (diese schneiden dann schlechter als beim ersten Mal ab). Dieser Effekt würde jedoch nicht auftreten, wenn man die Gruppe nicht nach Leistung, sondern nach Geschlecht trennen würde.

Obwohl das Phänomen der statistischen Regression seit Galton (1886), der den Ausdruck »Regression zum Mittelwert« prägte, und seit Thorndike (1924), der die artifizielle Korrelation zwischen Ausgangswert und Veränderung formal einfach zeigen konnte, bekannt ist, hat sich dieses Wissen in der angewandten Forschung kaum durchgesetzt. So ist es z. B. derzeit in der Suchtforschung gang und gäbe, zu behaupten, dass fast zwei Drittel der Suchtkranken innerhalb eines Jahres remittieren (z. B. De Bruijn et al. 2006 oder Klingemann & Carter Sobell 2007), ohne dass diese Forscher auch nur in Erwägung ziehen, dass sich diese Veränderungen zu einem erheblichen Anteil als Regressionsartefakte erklären lassen.

Eine einfache Möglichkeit, Regressionsartefakte zu kontrollieren, ist ein Design mit einer Versuchsgruppe und einer Kontrollgruppe zu wählen, weil man so Versuchseffekte gegen Regressionsartefakte abgrenzen kann. Absurd wird das Ganze allerdings dann, wenn die Veränderungen in der Kontrollgruppe als Placeboeffekte, Wartelisteneffekte oder Spontanremissionseffekte bezeichnet werden, ohne auch nur einen Gedanken darauf zu verschwenden, dass es sich dabei mit hoher Wahrscheinlichkeit primär um artifizielle Regressionseffekte handelt. Auch Behandlungseffekte und Präventionseffekte, die nicht gegen eine Kontrollbedingung getestet werden, sind mit großer Wahrscheinlichkeit primär als Regressionsartefakte zu interpretieren.

3.5.4 Prävalenzbias

Wenn man die Prävalenz (Pr)[6] von Merkmalen, wie z. B. Depression, dichotom (»vorhanden« vs. »nicht vorhanden«) erfassen möchte und dabei die Messung nicht völlig fehlerfrei erfolgt, so entsteht zwangsläufig ein systematischer Fehler, den man mathematisch aber leicht korrigieren kann, wenn man die Sensitivität (Se)[7] und Spezifität (Sp)[8] der Messmethode kennt. Wenn ein Screeningverfahren z. B. 12 % einer Population als depressiv ausweist (observed prevalence $Pr_O = 12\,\%$) und wenn die Sensitivität des Messverfahrens $Se = 90\,\%$ und die Spezifität $Sp = 90\,\%$ beträgt, dann beträgt die tatsächlich Prävalenz bloß $Pr_T = 2{,}5\,\%$, wie man mit einer Formel von Rogan & Gladen (1978) leicht nachrechnen kann.

$$Pr_T = (Pr_O + Sp - 1) / (Se + Sp - 1) = (0{,}12 + 0{,}90 - 1{,}00) / (0{,}90 + 0{,}90 - 1{,}00) = 0{,}02 / 0{,}80 = 0{,}025 = 2{,}5\,\%$$

7 Sensitivität = Prozentsatz der Merkmalsträger, die durch das Verfahren korrekt identifiziert werden

8 Spezifität = Prozentsatz der Nicht-Merkmalsträger, die durch das Verfahren korrekt identifiziert werden

Sensitivitäten und Spezifitäten um 90 % werden bei Screeningverfahren durchwegs als exzellent aufgefasst, auch wenn sie, wie gezeigt werden konnte, einen enormen systematischen Fehler produzieren. Der Prävalenzbias wird, obwohl er seit langem bekannt und leicht zu belegen ist, in der human- und sozialwissenschaftlichen Forschung so gut wie nie beachtet. Außer bei Wahlprognosen, wo Meinungsforschungsinstitute, um sich nicht völlig lächerlich zu machen, erfahrungsgestützte Korrekturformeln anwenden, werden bei Umfragen fast ausnahmslos die rohen Ergebnisse – mit oder ohne Konfidenzintervalle – präsentiert. Dieser systematische Fehler (Prävalenzbias) ist in den Extrembereichen – also bei sehr hohen und sehr niedrigen Prävalenzen – besonders groß, wird, wie alle systematischen Fehler auch bei großen Stichprobenumfängen nicht kleiner und wird von Konfidenzintervallen nicht miterfasst.

3.5.5 Der Umgang mit Signifikanzen

Es ist offensichtlich, dass es notwendig ist, beobachtete Unterschiede und Zusammenhänge zufallskritisch abzusichern – und zu diesem Zweck ist der Signifikanztest erfunden worden. Bei der Anwendung von Signifikanztests gibt es allerdings zahlreiche Probleme. Pointiert und nicht zu Unrecht beschrieb Sponsel (2006) den Signifikanztest als »*Szientismus zwischen numerologischer Esoterik, Gaukeln und Betrug*«.

Das logische Hauptproblem beim Signifikanztest ist, dass der Alphafehler die Wahrscheinlichkeit dafür beschreibt, mit der ein beobachteter oder noch größerer Unterschied auftreten würde, wenn die H_0-Annahme zutrifft, dass wir aber in der Praxis die umgekehrte Situation beurteilen möchten, nämlich wie groß ist die inverse Wahrscheinlichkeit angesichts der vorliegenden Daten, dass nämlich die H_0-Annahme zutrifft. Bei der Interpretation von Signifikanz wird in der Regel so getan, als ob der Schluss von der Theorie auf die Daten und der Schluss von den Daten zur Theorie gleichwertig sei – eine Unsinnigkeit, die Gigerenzer (1993) *"The Bayesian Id's wishful thinking"* nennt, da man die inverse Wahrscheinlichkeit mithilfe des Theorems von Bayes berechnen könnte, wenn man die unbedingte Wahrscheinlichkeit dafür, das H_0 zutrifft, kennen würde.

Sieht man von diesem grundlegenden Problem ab, so ist eine der zentralen Forderungen zur Signifikanzrechnung, dass man diese nur für Entscheidungsstudien (Hypothesenprüfung) einsetzen darf; d. h. wenn die Zusammenhänge aufbauend auf explorative Studien bereits relativ gut verstanden worden sind und sich die Entscheidung für oder gegen die Hypothese auf eine oder wenige Variablen stützt. Außer in wenigen Bereichen, wie z. B. bei Zulassungsstudien im Pharmabereich, wo die Zulassungsbehörde die Einhaltung fundamentaler statistischer Regeln erzwingen kann und dies auch tut, werden diese Grundsätze in den meisten Forschungsfeldern ignoriert. Der Großteil der empirischen Studien ist zwar eindeutig explorativ angelegt, aber es werden nichtsdestoweniger Signifikanzen gerechnet und dem Leser implizit als Entscheidungsstudien präsentiert. Meist werden dabei unzählige Signifikanztests gerechnet und nur die signifikanten Ergebnisse veröffentlicht, was von Statistikern als »Fischen nach Signifikanzen« konsequent abgelehnt wird.

4 Schlussfolgerungen

Es ist ohne Frage nicht leicht, die Wirksamkeit von Suchtpräventionsmaßnahmen wissenschaftlich exakt zu belegen, aber wir dürfen und können Suchtprävention deswegen nicht einfach aufgeben, so wie wir Erziehung, Schulunterricht, Psychotherapie, Wirtschaftspolitik, Umweltpolitik und vieles mehr nicht bleiben lassen, bloß weil es schwer ist, wissenschaftlich eindeutige Belege für die Sinnhaftigkeit einzelner Maßnahmen zu bringen.

Was wir in diesem Zusammenhang brauchen, ist allerdings eine kritische Grundhaltung, die die grundlegenden erkenntnistheoretischen, psychologischen und methodischen Grenzen zur Kenntnis nimmt und nicht einfach verleugnet. Wir brauchen die Bereitschaft und den Mut, grundlegende Begriffe zu definieren und unüberwindbare Erkenntnisgrenzen anzuerkennen. Wir brauchen die Konsequenz und Hartnäckigkeit, möglichst divergent nach konkurrierenden Hypothesen und Theorien zu suchen, und ausreichend Motivation, trotz vieler Schwierigkeiten nicht aufzugeben. Bei der Entwicklung von Hypothesen und Theorien müssen wir auf Commonsense, Forschungsergebnisse, Erfahrung und vor allem kritisch-rationales

Denken bauen. Wie Popper (1972) vorschlug, sollten wir unsere neuen Hypothesen und Theorien einige Zeit lang verteidigen, bis wir sie wirklich im Detail verstanden haben, und diese in Anschluss daran kompromisslos in Frage stellen, um ein höheres Niveau des Verstehens erreichen zu können.

Von zentraler Bedeutung ist hier Ambiguitätstoleranz. Es muss uns möglich sein zuzugeben, dass wir Phänomene ganz und gar nicht verstehen und dass wir in manchen Situationen große Zweifel haben, ohne unsere Kompetenz als Wissenschaftler und Evaluatoren generell in Frage zu stellen. Das ist keine leichte Forderung in einer Welt, wo Wissenschaftler sich und ihre Arbeit verkaufen müssen und dabei mit Kollegen in Konkurrenz stehen, die sich nicht scheuen, bereitwillig zu allen Themen Aufträge anzunehmen, als Experten für alles und jedes auftreten und nie Ungewissheiten im Raum stehen lassen.

Eine weitere wichtige Voraussetzung sind adäquate Arbeitsbedingungen, die es Praktikern, Evaluatoren und Forschern ermöglichen, sich lange genug mit einer Thematik zu beschäftigen, um diese grundlegend so weit zu verstehen, dass sie darauf aufbauend ihre Arbeit ständig weiterentwickeln können. Diese Personen brauchen auch genügend Zeit, um sich in Supervisionen, Intervisionen und Arbeitskreisen, sowie bei Fortbildungen und bei Tagungen mit Kollegen austauschen zu können. Unglücklicherweise geht der globale Trend derzeit in eine ganz andere Richtung – weg von Basisfinanzierung und »Zentren der Exzellenz« oder "Think tanks" hin zu knapp kalkulierten, zeitlich befristeten Projekten im Rahmen von drittmittelfinanzierter Auftragsforschung. Newton popularisierte einen Ausspruch von John of Salisbury in dessen "Metalogicon" (McGarry 1962), der gemeint hatte, dass Wissenschaftler mit Zwergen vergleichbar seien, die mehr als Riesen sehen könnten und einen weiteren Horizont hätten als diese, aber nicht, weil ihre Augen besser wären oder sie andere physiologische Vorteile hätten, sondern weil sie hochgehoben wurden und auf den Schultern von Riesen stehen. Wissenschaftswachstum im Sinne dieser klugen Metapher erfordert aber, dass Wissenschaftler Zeit haben, sich auf die Schultern von Riesen zu begeben, und nicht bloß vorgeben, auf deren Schultern zu stehen.

Ich kann aus meiner langen Erfahrung als Forscher sagen, dass die meisten Evaluatoren und Wissenschaftler großes Interesse daran haben, ihren Wissensstand zu vergrößern und echte Erkenntnis zu erlangen, aber hier gibt es eine dreistufige Hierarchie zu beachten. Von primärer Bedeutung ist es, an genügend lukrativen Projekten beteiligt zu sein, um davon leben zu können. Wer das nicht schafft, spielt bald nicht mehr mit in der Arena der Wissenschaft. An zweiter Stelle geht es darum, einen Ruf als verlässlicher Evaluator und Forscher zu begründen und zu verstärken, da man nur so an weiteren interessanten, in der Fachwelt beachteten Projekten beteiligt wird. An dritter Stelle kommt, das eigene Wissen und Verständnis zu erweitern. Da sich die meisten Forscher primär um Stufe eins und zwei kümmern müssen und immer weniger Zeit und Energie für Stufe drei bleibt, ist Erkenntnis für viele Evaluatoren und Forscher wie ein wertvolles Juwel, das sie zwar gerne besitzen würden, das sie sich aber nicht leisten können.

Literatur

Alvin DF (2007) Margins of Error. Wiley, Hoboken, New Jersey

Anderson P, Baumberg B (2006) Alcohol in Europe: A Public Health Perspective. Institute of Alcohol Studies, London

Babor T, Caetano R, Casswell S et al. (2005) Alkohol – Kein gewöhnliches Konsumgut Forschung und Alkoholpolitik. Hogrefe, Göttingen

Babor TF, McGovern T (2008) Dante's inferno: Seven deadly sins in scientific publishing and how to avoid them. In: Babor TF, Stenius K, Savva S et al. (eds) Publishing addiction science: a guide for the perplexed (2nd edition). International Society of Addiction Journal Editors, London

Beywl W (ed) (2003) Evaluation Standards (DeGEval-Standards). deutsche Gesellschaft für Evaluation, Köln

Bosbach G, Korff JJ (2011) Lügen mit Zahlen. Wie wir mit Statistiken manipuliert werden. Heyne, München

Bühringer G, Batra A (2004) Industriesponsoring: Teufelszeug, akzeptable, sinnvolle oder notwendige Finanzierungsquelle? Sucht 50 (2): 99–101

Clayton RR, Cattarello A (1991) Prevention Intervention Research: Challenges and Opportunities. In: Leukefeld CG, Bukoski WJ (eds) Drug Abuse Prevention Intervention Research: Methodological Issues. NIDA Research Monograph 107, Rockville

Cronbach LJ, Gleser GC, Nanda H et al. (1972) The Dependability of Behavioral Measurements: Theory of Generalizibility for Scores and Profiles, Chapter 1, The Multifacet Concept of Observational Procedures. John Wiley & Sons, USA

Curtis GN (2006) The fallacy files www.fallacyfiles.org (Stand: 31.06.2006)

Davis RM (1997) The Farmington consensus Statement on editorial guidelines for addiction Journals. Tobacco Control 6: 167–168
De Bruijn C, Van den Brink W, de Graaf R et al. (2006) The three Year Course of Alcohol Use Disorders in the General Population: DSM-IV, ICD-10 and the Craving Withdrawal Model. Addiction 101: 385–392
Diederich W (1978) Einleitung. In: Diederich W (Hrsg) Theorien der Wissenschaftsgeschichte – Beiträge zur diachronischen Wissenschaftstheorie. Suhrkamp, Frankfurt a.M., 7–51
Doll HA (2008) Making sense of the numbers: CER, EER, ARR, RR, RRR, NNT, CIs. The Foundation Years 5 (Supplement 1): 2009, Pages 1–7
Donabedian A (1980) The Definition of Quality and Approaches to its Assessment. Health Administration Press Ann Arbor, Michigan
Dörner D (1996) The Logic of Failure: Recognizing and Avoiding Error in Complex Situations, Perseus Books, Cambridge
Feyerabend P (1978) Der wissenschaftliche Realismus und die Autorität der Wissenschaften, Ausgewählte Schriften. Vieweg, Band 1, Braunschweig
Galton F (1886) Regression Towards Mediocrity in Hereditary Stature. The Journal of the Anthropological Institute of Great Britain and Ireland 15: 246–263
Gigerenzer G (1993) The Superego, the Ego, and the id in Statistical Reasoning. In: Kerren G & Lewis Ch. A Handbook for Data Analysis in the Behavioral Sciences. Lawrence Erlbaum, Hillsdale
Gigerenzer G (2002) Calculating Risks: How to Know When Numbers Deceive You. Simon & Schuster, New York
Gigerenzer G (2008) Bauchentscheidungen – die Intelligenz des Unbewussten und die Macht der Intuition. Goldmann, München
Hartnoll R (2004) Drugs and Drug Dependence: Lessons learned, challenges ahead. Council of Europe Publishing, Strasbourg
Hayek FA (1988) The Fatal Conceit. Routledge, Chicago
Hood Ch (1991) A Public Management for All Seasons?. Public Administration 69: 3–19
Hume D (1951) A Treatise of Human Nature, Book 3 of Morals, Part 1, Of Virtue and Vice in General. In: Watkins F (Ed) Hume Theory of Politics. Nelson and Sons, Edinburgh; Original: First Edition (1740) Thomas Longman, London
Huff D (1954) How to Lie With Statistics. Norton & Company, edition 1993, New York
Kahneman D (2012) Schnelles Denken, langsames Denken. Siedler, München
Klingemann H, Carter Sobell L (eds) (2007) Promoting Self-Change from Addictive Behaviors. Springer, New York
Kritz J, Lück HE, Heidbrink H (1990) Erkenntnis- und Wissenschaftstheorie. Leske, Opladen
Liessmann KP (2005) Evaluation: Wie viel wiegt Wissen?. Spectrum, 29.01.2005
Luhmann N (2000) Die Unwahrscheinlichkeit der Kommunikation. In: Pias C, Vogl J, Engell L et al. Kursbuch Medienkultur. Die maßgeblichen Theorien von Brecht bis Baudrillard. Deutsche Verlags-Anstalt, Stuttgart
McGarry DD (1962) The Metalogicon of John Salisbury. University of California Press, Berkeley
Moore GE (1903) Principia Ethica. Cambridge: Cambridge University Press (second paperback edition, 1960)
Pirie M (1985) The book of the fallacy. London: Routledge & Kegan Paul
Popper KR (1934) Logik der Forschung. J.C.B.Mohr (6. verbesserte Auflage, 1976), Tübingen
Popper KR (1972) Objective Knowledge – An Evolutionary Approach. Oxford University Press, New York
Reiter C (2005) Lesekompetenz von österreichischen SchülerInnen – ein internationaler Vergleich. Büchereiperspektiven 3: 18–21
Rogan WJ, Gladen B (1978) Estimating Prevalence from the Results of a Screening Test. American Jounnal of Epidemiology 107 (1): 71–76
Rothman KJ (1993) Conflict of Interest. The new McCarthyism in Science. JAMA 269: 2782–2784
Sackett DL, Rosenberg WMC, Gray M et al. (1996) Editorial: Evidence Based Medicine: What it is and what it isn't. British Medical Journal 312: 71–72
http://www.ncope.org/summit/pdf/Footnote1.pdf
Schmutterer I, Uhl A, Strizek J et al. (2008) ESPAD Austria 2007: Europäische SchülerInnenstudie zu Alkohol und anderen Drogen – Band 2: Validierungsstudie. Bundesministerium für Gesundheit, Wien
http://www.api.or.at/akis/download/espad%202007band2.pdf
Scriven M (1991) Evaluation Thesaurus, 4th Edition. Sage, Newbury Park
Scriven M. The Methodology of Evaluation. In: Tyler RW, Gagne RM, Scriven M (eds) (1967) Perspectives of Curriculum Evaluation. Rand-Mc. Nally, Chicago
Sponsel R (2006) Beweis und beweisen in der Statistik. Blicke über den Zaun zum Auftakt für eine integrative psychologisch-psychotherapeutische Beweislehre, Internet Publikation für Allgemeine und Integrative Psychotherapie www.sgipt.org/wisms/gb/beweis/b_stat.htm (Stand: 11.11.2006)
Steward I (1989) Does God Play Dice? The Mathematics of Chaos. Penguin, London
Thorndike EL (1924) The Influence Of The Chance Imperfections Of Measures Upon The Relation Of Initial Score To Gain Or Loss. Journal of Experimental Psychology 7: 225–232
Tukey JW (1977) Exploratory Data Analysis. Addison-Wesley, Reading
Uhl A (1998) Evaluation of Primary Prevention in the Field of Illicit Drugs: Definitions – Concepts – Problems. In: Springer A, Uhl A (eds) Evaluation Research in Regard to Primary Prevention of Drug Abuse. A COST-A6 Publication. European Commission Social Sciences, Brussels
www.api.or.at/lbi/download.htm
Uhl A (2000a) The Limits of Evaluation. In: Neaman R, Nilson M, Solberg U. Evaluation – A Key Tool for Improving Drug Prevention. EMCDDA Scientific Monograph Series, No 5, Lisbon
http://www.emcdda.europa.eu/html.cfm/index34013EN.html
Uhl A (2000b) Evaluation. In: Stimmer F (Hrsg) Suchtlexikon. Oldenbourg, München
Uhl A (2000c) Evaluation vs. Evalopathy: Support for Practical Improvement vs. Irrational Nuisance. In: Abstracts of

the 3rd Nordic Health Promotion Research Conference, Tampere, 6–9 September 2000. University of Tampere, Tampere

Uhl A (2002) Schutzfaktoren und Risikofaktoren in der Suchtprophylaxe. In: Röhrle B (Hrsg) Prävention und Gesundheitsförderung Bd.II. DGVT, Tübingen

Uhl A (2007) How to Camouflage Ethical Questions in Addiction Research. In: Fountain J, Korf DJ (eds) Drugs in Society European Perspectives. Radcliffe, Oxford http://www.radcliffe-oxford.com/books/samplechapter/0932/Chapt10-4e4c680rdz.pdf

Uhl A (2010) Evaluation of the Drug Prevention Activities: Theory. In: Uhl A, Ives R, Members of the Pompidou Group Prevention Platform (eds.) Evaluation of Drug Prevention Activities: Theory and Practice (2010), Council of Europe, Strasbourg

Uhl A, Bachmayer S, Kobrna U (2009a) Chaos um die Raucherzahlen in Österreich. Wiener Medizinische Wochenschrift, 4–13

Uhl A, Bachmayer S, Kobrna U et al. (2009b) Handbuch: Alkohol – Österreich: Zahlen, Daten, Fakten, Trends 2009. 3. überarbeitete Auflage. BMG, Wien http://www.api.or.at/sp/download/habaoe%2020009a.pdf

Uhl A, Strizek J, Puhm A et al. (2009c) Österreichweite Repräsentativerhebung zu Substanzgebrauch 2008, Band 1: Forschungsbericht. Bundesministerium für Gesundheit, Wien http://www.api.or.at/akis/download/gps2008band1.pdf

WHO (1946) The WHO Constitution. World Health Organization, New York

WHO (1986a) Ottawa Charter. World Health Organization, Geneva

WHO (1986b) Health Promotion: A Discussion Document on the Concepts and Principles. Health Promotion 1: 73–76 http://www.who.int/healthpromotion/Milestones_Health_Promotion_05022010.pdf

Wiener Krankenanstaltenverbund (2012) New Public Management http://www.wienkav.at/kav/gd/texte_anzeigen.asp?id=67 (Stand: 22.01.2012)

Mögliche und machbare Evaluationsdesigns – Gedanken zur Evaluation oder: von Kanonenkugeln und Köchen[1]

Thomas Elkeles

1 Einleitung

Die mir gestellte Frage, welche Evaluationsdesigns bei der Gesundheitsförderung für Erwerbslose möglich und machbar sind, deutet bereits an, dass es zwischen theoretischer Möglichkeit eines Studiendesigns und der Durchführbarkeit eines (angemessenen) Designs Unterschiede, wenn nicht gar Widersprüche geben kann. Mit der Themenstellung ist gleichzeitig umrissen, dass sich die Designs in einem Spektrum bewegen werden, aus dem unter verschiedenen Gesichtspunkten eine Auswahl zu treffen ist oder auch Adaptationen vorzunehmen sein werden.

Es ist auch nicht allzu viel vorweggenommen, wenn an dieser Stelle bereits festgehalten sei, dass eine solche Wahl vom konkreten Gegenstand und dessen Kontextbedingungen abhängig sein wird. Selbst wer der Auffassung ist, in der Evaluation sei stets ein bestimmter Goldstandard – wie der Untersuchungs-/Kontroll-Gruppen-Vergleich oder gar die Randomisierung – zugrunde zu legen oder zumindest anzustreben (dazu jedoch Grundsätzliches weiter unten in Abschnitt 2.2), wird sich meines Erachtens jedenfalls zumindest darauf einlassen müssen, wie Evaluationsgegenstand und Kontext mit solchen Idealen jeweils in Einklang gebracht werden können.

Bevor jedoch Gegenstand und Kontext von Interventionen zur Gesundheitsförderung von Arbeitslosen hinsichtlich der Frage nach verschiedenen Evaluationsdesigns hier näher betrachtet werden (Abschnitt 3), sollen zunächst einmal einige begriffliche Erläuterungen zur Evaluation und zur Evaluationsforschung sowie zu ihrer wissenschaftstheoretischen Verortung im folgenden Abschnitt vorgenommen werden.

2 Begriff und wissenschaftstheoretische Verortung von Evaluation

2.1 Evaluation

Es kann hilfreich sein, Erfahrungen aus der Alltagsbeobachtung und dem der Alltagssprache zugrunde liegenden Verständnis einer an bestimmten Maßstäben vorgenommenen Bewertung heranzuziehen, wenn von Evaluation die Rede ist. Wenn wir jedoch von vornherein Evaluation auf die wissenschaftliche Evaluation eingrenzen, bringt das den Vorteil mit sich, dem inflationären Sprachgebrauch entgegenzuwirken, der sich im Zusammenhang mit dem Begriff Evaluation seit einiger Zeit eingebürgert hat.

Eine wissenschaftliche Definition zielt darauf ab, von Evaluation in dem Sinne zu sprechen, dass spezifische Sachverhalte in einem objektivierten Verfahren und nach explizit auf den Sachverhalt bezogenen und begründeten Kriterien durch Personen bewertet werden, die zu dieser Bewertung in besonderer Weise befähigt sind (Kromrey 2001; Elkeles, Beck 2010). Sie ist eine methodisch kontrollierte, verwertungs- und bewertungsorientierte Form des Sammelns und Bewertens von Informationen. Empirisch-wissenschaftliche Forschung wird zur Evaluation, wenn sie einer intersubjektiv geltenden (normativen) Bewertung eines Sachverhaltes dient, die in einem objektivierten Verfahren und anhand explizit gemachter Kriterien und Maßstäbe vorgenommen wird.

Im Vergleich zu anderweitiger empirischer angewandter (Sozial-) Forschung hat sie weniger oder kaum Besonderheiten in der Methodik als vielmehr in dem spezifischen Erkenntnis- und Verwertungsinteresse.

Ein Großteil der Literatur erscheint unter dem Begriff Programmevaluation. Unter Programmen sollen komplexe Handlungsmodelle verstanden werden, die

- auf die Erreichung bestimmter Ziele gerichtet sind,

1 Beitrag ursprünglich erschienen in: Bellwinkel M, Kirschner W (Hrsg) (2011) Evaluation von Projekten der Gesundheitsförderung von Arbeitslosen. Reihe Gesundheitsförderung und Selbsthilfe (Hrsg: Bundesverband der Betriebskrankenkassen), Nr. 25, Bremerhaven, Wirtschaftsverlag NW, S. 31–51

- auf bestimmten, den Zielen angemessen erscheinenden Handlungsstrategien beruhen, und für deren Abwicklung
- finanzielle, personelle und sonstige Ressourcen bereitgestellt werden (Hellstern, Wollmann 1984).

Da wir es mit sozialen Programmen zu tun haben oder zumindest mit den sozialen Prozessen in (Interventions)Programmen, sei hier noch mit Rossi et al. (1988) ergänzt, dass es sich bei Evaluation um eine »systematische Anwendung sozialwissenschaftlicher Forschungsmethoden zur Beurteilung der Konzeption, Ausgestaltung, Umsetzung und des Nutzens sozialer Interventionsprogramme« handelt (Rossi et al. 1988).

In der Systematik von Rossi et al. (1988) werden drei Haupttypen von Evaluationsstudien unterschieden:

- Analysen zur Programmentwicklung, einschließlich Konzeptualisierung und Ausarbeitung einer geplanten Intervention.

Ein Beispiel könnte die Entwicklung und Implementierung eines psychosozialen Dienstes für bestimmte Zielpopulationen sein.

- Begleitforschung oder Monitoring als laufende Überwachung der Umsetzung und Überwachung eines Programms; Prozessevaluierung.

Ein Beispiel könnte die Dokumentation und prozessbezogene Auswertung eines Modellversuchs zur Einführung einer Fördermaßnahme für bestimmte Zielpopulationen sein.

- Abschätzung von Programmwirkungen und Programmnutzen; Ergebnis-, Nutzenevaluierung.

Ein Beispiel könnte die outcome-Bewertung einer bestimmten Intervention bei bestimmten Zielpopulationen sein.

Auf viele andere mögliche Unterscheidungen und deren Hintergründe (wie etwa die von formativer und summativer Evaluation) soll hier bewusst verzichtet werden.

Stattdessen wollen wir uns hier darauf konzentrieren, was entscheidende Fragen bei der Konstruktion bzw. Wahl eines Evaluationsdesigns für soziale Programme sind. Diese gelten bei der Gesundheitsförderung von Erwerbslosen im Speziellen wie auch generell bei sozialen Programmen.

Statt etwa ein Kochbuch der Evaluation zu unterbreiten, dessen Anwendbarkeit von der Art und Anzahl der verfügbaren Kochtöpfe, -geräte und -utensilien abhängig sei, soll im Folgenden unter Rekurs auf eine manifestartige Darstellung der »Realistischen Evaluation« durch die beiden britischen Soziologen Ray Pawson und Nick Tilley deren Ansatz umrissen werden, wie mit der zentral mit dem Begriff der Evaluation verbundenen Frage der Kausation umzugehen sei. Alle denkbaren Evaluationsdesigns sind implizit oder explizit Antworten auf die – allerdings nicht stets wirklich oder wirklich angemessen reflektierte – Frage, wie ein Experiment konstruiert sein muss, damit Effektivität und Effizienz eines sozialen Interventionsprogramms kausal nachweisbar werden.

2.2 Machte das Programm die Subjekte gesünder, reicher, sicherer?

Diese Frage stellen Pawson und Tilley (1997) in ihrem Text – sie ist hier in Anlehnung an deren streckenweise betont (britisch?) humorvolle Schreibweise zur Zwischenüberschrift erhoben.[2] Anders formuliert heißt die Frage, ob es wirklich das Programm – und nur das Programm – war, welches verantwortlich für die beobachteten Veränderungen bei den untersuchten Subjekten war – und zwar in jener üblichen Art positivistischer experimenteller Evaluation, wie sie nach Darlegung der Autoren jedoch unangemessen für den Charakter sozialer Erklärung sei (ebda., S. 31).

Die Frage nach der Kausalität und ihrem Nachweis zieht sich als Disput durch Jahrhunderte von Philosophiegeschichte. Pawson und Tilley (1997) bringen sie, in Anknüpfung an Harré (1972), auf die Unterscheidung zwischen sukzessionistischer und generativer Kausationstheorie. Diese beiden Rekonstruktionen würden vieles von dem zusammenfassen, was bei unserem Verständnis von Kausalität im Disput stehe. Beide dieser

2 Original-Kapitelüberschriften bei Pawson und Tilley (2010) lauten z. B. (Übersetzung d. Verf.): »Eine Geschichte der Evaluation in 28 und einer halben Seite« oder »No Smoking ohne Feuermechanismen: eine ›Realistische‹ Konsultation«.

großen Metatheorien gingen von der Annahme aus, Kausalerklärung in der Wissenschaft sei eine Angelegenheit des Besitzes einer Methode, die jene Fälle, in denen X mit Y auf regelmäßige gesetzmäßige Weise verknüpft ist, von jenen unterscheiden könne, in denen die Assoziationen durch Zufall zustande kommen. Der strittige Punkt zwischen den beiden Perspektiven sei, wie man starke Evidenz erbringe, um solch einen Anspruch zu unterstützen.

Hierbei folgten Sukzessionisten der Ansicht von David Hume (1989), Kausation sei unbeobachtbar und man könne sie lediglich auf der Basis von Beobachtungsdaten ableiten. Es gehe also darum, eine kontrollierte Folge von Beobachtungen einzurichten, welche die Kausalbeziehungen von vorgetäuschten Assoziationsbeziehungen unterscheiden lassen. Der logische Bezugsrahmen hinter diesem Beobachtungsansatz sei durch die berühmten Methoden von John Stuart Mill (1961) etabliert worden. Am Ende dieser jahrhundertelang zurückreichenden Auffassungen habe dann das Quasi-Experiment und das ›klassische experimentelle Design‹ in der Evaluation gestanden (vgl. Tab. 1).

Tabelle 1
Das klassische experimentelle Untersuchungsdesign
Quelle: Eigene Darstellung nach Pawson u. Tilley 2004, S. 32

	Pretest	Treatment	Posttest
Untersuchungsgruppe	O_1	X	O_2
Kontrollgruppe	O_1		O_2

Das gesamte forschungsseitige Regime von Manipulation, Kontrolle und Beobachtung brauche im OXO-Modell keinerlei weiteren Informationen, um festzustellen, dass Treatment und Outcome auf eine Weise miteinander verbunden seien, dass der einzige Unterschied in der Applikation der Initiative liege und jegliche Unterschiede im Verhalten zwischen den Gruppen von dieser Applikation rührten. Dieser sukzessionistische Kausalitätsbegriff entspreche der erkenntnistheoretischen Tradition, dass Kausalkräfte selbst nicht beobachtet werden könnten und Kausalität extern sei.

Die generative Kausationstheorie hingegen gehe davon aus, dass Kausation sowohl extern als auch intern sei. Deren Experimente zielten auch auf regelhafte und systematische Muster von inputs und outputs, Gründen und Effekten, etablierten aber ihre Verbindung auf deutlich unterschiedliche Weise.

Vor näherer Erläuterung seien zwei an unterschiedlichen Stellen sich findende Anmerkungen von Pawson und Tilley (1997) erwähnt. Zum einen findet sich ein erläuternder Exkurs, bereits in der Alltagssprache mache man Gebrauch davon, dass wir beobachtbaren Veränderungen (ein Fall, ein System, ein Ding, eine Person befänden sich in Transition, vgl. ebda., S. 33) nicht ausschließlich externen, sondern auch internen Kräfte zuschreiben.

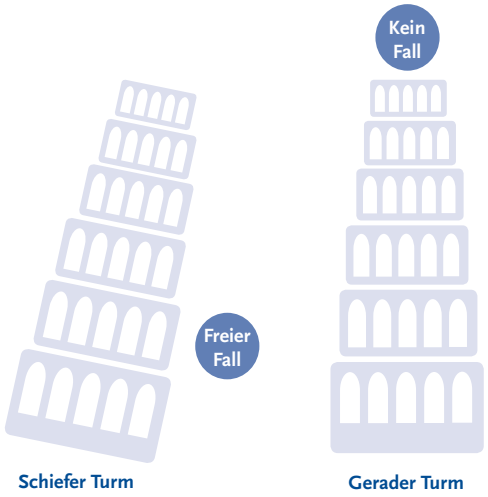
So mögen wir berichten, Schießpulver sei explodiert, die Ökonomie sei in Rezession übergegangen, der Gefangene sei rehabilitiert worden. Bei der Erklärung dieser Transitionen verweisen wir auf den extern beobachtbaren Grund (wie einen Zündfunken, eine Ölkrise, ein Erziehungslager), bauen aber als Teil der Erklärung auch auf irgendeinen internen Befund dessen, was sich geändert hat (wie etwa Zustand oder chemische Zusammensetzung des Schießpulvers, Struktur und Entwicklungsniveau der Ökonomie, den Charakter oder die Disposition des Gefangenen). Diese internen Kräfte schienen uns sowohl in Alltags- als auch in wissenschaftlicher Erklärung wichtig, denn sie erlaubten uns eine Bedeutungsbestimmung der Gelegenheiten, bei denen die Kausalbeziehung abwesend ist (wenn etwa der Funke das nicht zusammengepresste Schießpulver nicht entzündet oder wenn eine Ölkrise ölproduzierende oder Subsistenz-Ökonomien unangegriffen lässt oder wenn das Erziehungslager die ohnehin schon hartgesottenen Missetäter lediglich stählt).

Zum anderen seien zwei – vermutlich hypothetische – Beispiele erwähnt, die Pawson und Tilley (1997) als Antworten auf die von ihnen selbst als provokativ exkulpierte Frage geben, wie viele Experimente in den Naturwissenschaften wohl der Logik von Untersuchungs- und Kontrollgruppen folgten.

So fragen sie, ob wir etwa den Einfluss der Erdanziehung auf einen fallenden Körper verstehen, wenn wir die Bewegung einer von einem schiefen Turm herabfallenden Kanonenkugel beobachten und sie dabei mit der Bewegung einer anderen

Kanonenkugel vergleichen, die auf einer Turmspitze liegen bleibt (Abb. 1).[3] Und verstehen wir das Verhalten von Atomen, die in einem besonderen Beschleuniger zertrümmert werden, indem wir sie mit den Aktivitäten solcher vergleichen, die nicht diese Behandlung erfahren? Man müsse kein Wissenschaftshistoriker sein, um diese Fragen zu beantworten.

Abbildung 1
Kanonenkugel-Experiment – Untersuchungs- und Kontroll-Design



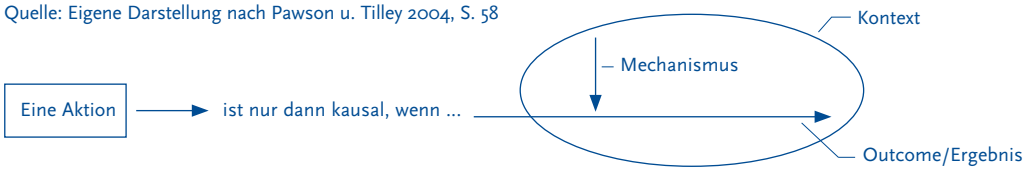
»Experimente in den Naturwissenschaften tendieren zur generativen Logik, während sozialwissenschaftliche Experimente hauptsächlich der sukzessionistischen Logik folgen« (ebda., S. 57).

Das in der Physik stets angewandte Erklärungsformat berücksichtige, was Axiom realistischer Erklärungen sei: Kausale outcomes resultieren aus Mechanismen, welche in Kontexten wirken (vgl. Abb. 2).[4]

In diesem MKO-Modell der generativen Kausation werden outcomes durch die Aktion partikularer Mechanismen in partikularen Kontexten erklärt, und diese Erklärungsstruktur wird über die Zeit durch eine Kombination von Theorie und experimenteller Beobachtung etabliert.

Während es die Logik des OXO-Modells ist, eine Situation so zu kontrollieren, dass alleinig ein Wechsel beim Treatment verantwortlich für die beobachteten outcomes sein kann, arbeite man beim MKO-Modell nicht mit einer derart eindimensionalen Logik. Hier habe ein Experiment den zu untersuchenden Mechanismus so zu triggern, dass sichergestellt ist, dass dieser aktiv ist, und zweitens müsse jede Beeinflussung mit der Operation des Mechanismus verhindert werden. Dies könne mit Bhaskar (1975) bezeichnet werden als ›experimentelle Produktion‹ und ›experimentelle Kontrolle‹. Aufgabe des Experimentalisten sei es, das gesamte Experimentalsystem zu manipulieren,

Abbildung 2
Generative Kausation
Quelle: Eigene Darstellung nach Pawson u. Tilley 2004, S. 58



... ihr Outcome durch einen Mechanismus getriggert wird, der in einem Kontext wirkt.

3 Das Kanonenkugeln-Experiment in der Physik mag auch deshalb hypothetisch sein, weil – wie der Physik-Leistungskursschüler Daniel Elkeles, geb. 1994, hierzu bei der Abfassung dieses Beitrags sogleich spontan angemerkt hat – auch auf die oben liegenbleibende Kugel die Erdanziehungskraft wirkt, ansonsten würde sie vielmehr über dem Turm schweben. Man mag daher annehmen, dass Physiker auch aus diesem Grunde in Wirklichkeit nicht zu diesem Experimentdesign greifen würden. Gleichwohl bleibt das Beispiel illustrativ für die Bevorzugung eines bestimmten Untersuchungsdesigns bei Fehlen dessen sinnvoller Voraussetzungen, wie sie idealtypisch in der Arzneimittelforschung vorliegen.

4 Pawson und Tilley (1997) führen dies unter anderem an Bernoulli und den Gasgesetzen wie auch anhand von Alltagswissen aus, wie: Es ist uns bekannt, dass Schießpulver nicht stets zur Entzündung durch eine Flamme kommt. So wissen wir, dass es keine Explosion geben wird, wenn die Bedingungen hierfür nicht richtig sind, etwa wenn die Mixtur feucht ist, zu wenig Pulver vorhanden ist, kein Sauerstoff anwesend ist, die Dauer der Hitzeanwendung zu kurz ist etc. (ebda., S. 58). Schließlich hätten wir das auch schon in Filmen öfters beobachtet.

d. h. die gewünschten Beziehungen zwischen unabhängiger und abhängiger Variable selbst zu produzieren. Der generative Experimentalist sei also ein Systembildner und die entscheidende Evidenz werde nicht durch kontrollierte Beobachtung, sondern (insofern, d. V.) durch Arbeit gebildet. Ähnliches geschehe in der Physik, wo Kontrolle nichts mit Kontrollgruppen zu tun habe, sondern die Form von ›Laborkontrolle‹ annehme (ebda., S. 59f.).

Fokussiert auf soziale Realität, soziale Kausation und sozialen Wandel (ebda., S. 63ff.), gehe es bei der kausalen Erklärung von Programmsystemen darum, folgende fünf Hauptpunkte zu berücksichtigen:

- Eingebundenheit (von menschlichem Handeln in einen breiten Rahmen sozialer Prozesse)
- Mechanismen
- Kontexte
- Gesetzmäßigkeiten und
- Wandel.

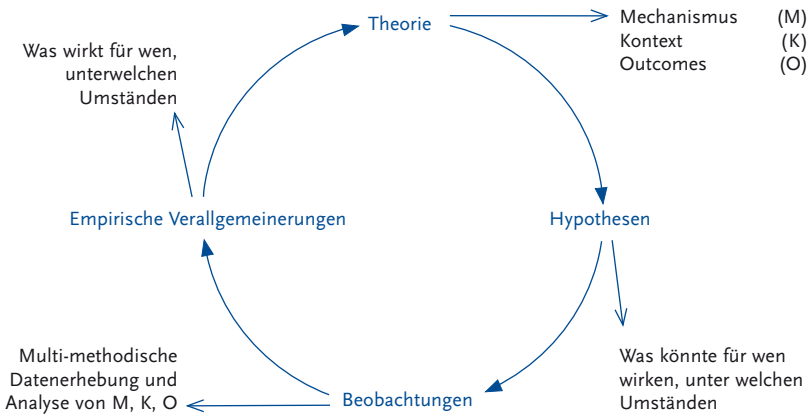
Nicht ein fixes Design führe hierzu (d. Verf.), sondern die Berücksichtigung dieser »Ingredienzien« anhand zweier Fragen (ebda., S. 75ff.):

1) Welches sind die Mechanismen für Wandel, die von einem Programm getriggert werden und wie wirken diese bestehenden sozialen Prozessen entgegen?
2) Welches sind die sozial und kulturell notwendigen Bedingungen, die verändert werden müssen, um Mechanismen zum Wirken zu bringen, und wie sind diese in und zwischen Programmkontexten verteilt?

Die Erläuterung des Designs realistischer Evaluation kann hier beschränkt werden auf die Darstellung des Evaluationszirkels (vgl. Abb. 3).

Auch die Konstruktion realistischer Daten, wie sie Pawson und Tilley (1997) vertreten, soll hier nicht näher ausgeführt werden. Selbstverständlich wird man sich aller dem Gegenstand angemessener Datenarten bedienen, und es gibt hierbei keinerlei grundsätzlichen Unterschied etwa zwischen quantitativen und qualitativen Daten und Methoden. Es sei nur abschließend darauf hingewiesen, dass Pawson und Tilley (1997) im entsprechenden Kapitel 6 ihres Buches ein gemeinsames Defizit im Methodenstreit fokussieren. Sowohl die Ansätze der ›dogmatischen Puristen‹ wie auch der ›pragmatischen Pluralisten‹ seien alle durchweg »datengeleitet« (data-driven strategies). Je nach bevorzugter Methode bestehe die Aufgabe dann darin, die Gedanken und Handlungen der Subjekte zu bestimmen. Die Datenerhebungsmethoden seien daher alle unter der Arbeitshypothese konstruiert, dass das Subjekt und der im Interview behandelte Gegenstand ein und dasselbe seien. Demgegenüber sei im realistischen (theory-driven) Modell die Theorie des Forschers, der Untersuchungsgegenstand des Interviews und das Subjekt (der stakeholder)

Abbildung 3
Der Realistische Evaluationszyklus
Quelle: Eigene Darstellung nach Pawson u. Tilley 2004, S. 85



dazu da, diese Theorie zu bestätigen, zu falsifizieren und sie vor allem zu verfeinern (ebda., S. 155).

Weitere Ausführungen und auch praktische Beispiele sind diesem Buch zu entnehmen, wir wollen nun jedoch zurückkehren bzw. voranschreiten zur Frage der Designs bei der Evaluation von Programmen zur Gesundheitsförderung von Erwerbslosen.

3 Gesundheitsförderung für und bei Erwerbslosen

In einem Gutachten für den BKK-Bundesverband haben wir den Kenntnisstand zu Arbeitslosigkeit und Gesundheit und zur Intervention mittels Gesundheitsförderung und Gesundheitsmanagement untersucht (Elkeles, Kirschner 2004). An den damaligen Befunden (Evidenz von Assoziationen bei häufig nur gering ausgeprägten Kenntnissen zur Kausalität und einer schwachen und teilweise konfligierenden Befundlage zum Erfolg von Interventionen) hat sich seitdem im Großen und Ganzen nur insoweit etwas verändert, als nunmehr einige Ergebnisevaluationen von Interventionen in Deutschland vorliegen (Kirschner, Elkeles 2006).[5]

Der Stand von Theorie und von Kenntnissen über soziale, psychosoziale und gesundheitliche Wirkmechanismen und deren Faktoren – zentrale Bestandteile der Forderungen seitens der Realistischen Evaluation – bleibt jedoch nach wie vor unbefriedigend.[6]

So gibt es bereits grundsätzlich zwei Interpretationsmöglichkeiten der Richtung von zugrunde liegenden Ursache-Wirkungs-Zusammenhängen für die schlechtere Gesundheit von Arbeitslosen: zum einen die Annahme von Kausaleffekten (der Arbeitslosigkeitsstatus führt zu gesundheitlichen Belastungen) und zum anderen die Annahme von Selektionseffekten (die schlechtere Gesundheit von Arbeitslosen ist durch häufigere und längere Arbeitslosigkeit von primär gesundheitlich Belasteten bedingt; Kausations- und Selektionshypothese, vgl. Elkeles, Seifert 1992; 1993a–c; 1996; Elkeles, Bormann 1999; 2002; Elkeles 2001; 2003). Eine Metaanalyse von Quer- wie auch Längsschnittstudien stützte anhand der Längsschnittstudien sowohl die Kausations- als auch die Selektionshypothese, wobei die Verursachung von psychischen Symptomen durch Arbeitslosigkeit der wichtigste kausale Faktor sei (vgl. Paul et al. 2006).

Jedenfalls zur Selektionshypothese wurden bisher eher sozialepidemiologische Interpretationsmuster statt gesellschaftstheoretischer Theorien benutzt, zu denen z. B. auch der Verweis auf den bekannten healthy worker effect gehört (in Populationen von erwerbstätig Beschäftigten findet sich ein durchschnittlich besserer Gesundheitsstatus).

Hinsichtlich der Annahme subjektiver Folgen von Arbeitslosigkeit für die psychische und/oder körperliche Gesundheit (Kausationshypothese) wurden bisher etwa stresstheoretische Erklärungen, kritische Lebensereignisse, die Theorie der erlernten Hilflosigkeit (Seligman 1979) und die Übertragung der Stigma-Theorie (Goffman 1975) erwogen, und häufig wird auf die Deprivationstheorie von Marie Jahoda Bezug genommen (Jahoda et al. 1975; vgl. Kieselbach, Beelmann 2006). Hinsichtlich des Einwandes, dass Jahodas Deprivationstheorie als Idealisierung der Erwerbsarbeit verstanden werden kann, scheinen auch die in der Arbeits- und Industriesoziologie reflektierten neueren Entwicklungen in der Arbeitswelt mit ihren Chancen und Risiken (vgl. Kratzer 2006) bedeutenswert für eine Theorie von ›Arbeitslosigkeit und Gesundheit‹ bzw. deren Weiterentwicklung. Im Zuge dieser Entwicklung entgrenze sich auch die Dichotomie Erwerbstätigkeit/Erwerbslosigkeit einerseits wie auch eine solche zwischen der Arbeitswelt mit ihren psychosozialen und gesundheitlichen Ressourcen und der Arbeitslosigkeit mit ihren psycho-sozialen und gesundheitlichen Risiken andererseits.

Damit mittels Evaluation die Theorie zur Kausalität von Veränderungen bei entsprechenden Interventionen getestet und weiterentwickelt werden kann, ist es erforderlich, eine solche Theorie überhaupt zugrunde zu legen, was prima vista nach wie vor ein deutliches Manko auch neuerlicher Interventionen und Evaluationen zur Gesundheits-

5 vergl. Bellwinkel, Kirschner 2011

6 Was jedoch an Mechanismen bekannt ist oder zu sein scheint, wird in allen Literaturübersichten (vgl. z. B. Kieselbach, Beelmann 2006; Mohr 2010) durch Moderatorvariablen modifiziert, zu denen u. a. das Ausmaß sozialer Absicherung bei Erwerbslosigkeit und die Höhe der Arbeitslosenquote und die kollektiven Erfahrungen mit Erwerbslosigkeit gehören. »Das nationale, regionale oder lokale Ausmaß von Arbeitslosigkeit hat Einfluss auf die Wahrnehmung ihrer Problemhaftigkeit. Hohe lokale Arbeitslosigkeit kann zu geringerem individuellen Stresserleben führen, gleichzeitig aber auch den gegenteiligen Effekt haben« (Elkeles, Kirschner 2004). Wie bei den Physik- und Chemiebeispielen von Pawson und Tilley (1997) ist also bekannt, dass ein Mechanismus einerseits innerer, andererseits Kontext-Bedingungen bedarf, welche es in der Evaluation angemessen zu berücksichtigen gilt.

förderung von Erwerbslosen und damit ein erstes *memento* dieses Beitrags ist.

3.1 Interventionsansätze

Programme zur gesundheitlich-sozialen Intervention bei Arbeitslosen lassen sich als Ansätze reaktiver Prävention kennzeichnen, während proaktive Interventionen sich auf den Arbeitsmarkt selbst zu richten haben bzw. hätten. Vor dem Hintergrund dieser Unterscheidung sind Programme und Maßnahmen der Prävention und Gesundheitsförderung für Arbeitslose kompensatorischen Charakters. Sie sind bereits deshalb, wenn nicht sogar vor allem deshalb, sinnvoll, um zum Erhalt und zur Förderung der Arbeitsfähigkeit und damit der Wiederbeschäftigungschancen beizutragen.

Dabei ist zunächst die unterschiedliche Bedarfslage von (noch) nicht und bereits erkrankten Arbeitslosen zu berücksichtigen. Für letztere dürften Maßnahmen der Gesundheitsförderung, wie sie üblicherweise von Krankenkassen im Rahmen des § 20 SGB V eingesetzt werden (überwiegend Individualprävention mit verhaltenspräventiven Maßnahmen), nicht ausreichend sein. Wir haben daher den Bedarf für Erwerbslose mit teilweise bereits massiven Einschränkungen als Gesundheitsmanagement bezeichnet (Elkeles, Kirschner 2004; Kirschner, Elkeles 2006). Dieses ist im Rahmen der AmigA-Projekte aufgegriffen worden.[7]

Der aktuelle Entwicklungsstand in der Diskussion und Praxis der Gesundheitsförderung bei Arbeitslosen ist neben einer

a) zielgruppenadäquaten Ausrichtung[8]
b) durch kombinierte Strategien der Gesundheits- und Beschäftigungsförderung gekennzeichnet (Arbeitsmarktintegrative Gesundheitsförderung).[9]

Gegenüber der Gesamtbevölkerung oder anderen sozialen Gruppen und Settings weisen nun allerdings Arbeitslose einige Besonderheiten in Zielgruppe, Identifikation, Erreichbarkeit und Programmakzeptanz auf. Sie sind eine heterogene Zielgruppe mit geringem Organisationsgrad, hoher Fluktuation, hoher sozialer Differenzierung und (partieller) Stigmatisierung ohne regelmäßige institutionelle Erreichbarkeit wie in Betrieben oder in Schulen.

Würde also das überwiegend bestehende Spektrum an »Gesundheitsförderungs«-Kursen verstärkt an Arbeitslose herangetragen, so bestünde die Gefahr, dass bei der Integration von Gesundheitsberatung und Gesundheitsförderung in die Sozialtechnologie des Profilings und Fallmanagements im Rahmen der Beschäftigungsförderung bei den Job-Centern (vgl. Elkeles, Michel-Schwartze 2008) Gesundheitsförderung wiederum auf verhaltensverändernde Strategien der Individualprävention verkürzt wird.

Besonders im Umfeld des beschäftigungsorientierten Fallmanagements bei Job-Centern könnte sich ein Grundkonflikt des gesellschaftlichen Umgangs mit Gesundheit verschärft stellen, wie er allerdings durch die Berücksichtigung der Ottawa-Charta ausgeschlossen sein sollte: nämlich die Frage, ob Gesundheit ein Recht oder eine Pflicht sei.

Bei nicht wenigen derjenigen von Elkeles und Kirschner (2004) untersuchten Projekte, die im behördlichen Kontext des Arbeits- oder Sozialamts stattgefunden hatten, fand sich ein Zwang zur Teilnahme an Angeboten, da ansonsten Sanktionen, also Leistungseinschränkungen, nicht ausgeschlossen waren.

Damit aber wäre eine Pflicht zur Gesundheitsberatung und ggf. Gesundheitsförderung etabliert, welche nicht nur dem sozialemanzipativen Charakter von Gesundheitsförderung widerspricht, sondern vielmehr auch grundsätzliche Voraussetzungen deren Wirksamkeit untergräbt. Aus der Ottawa-Charta ist damit für die Gesundheitsförderung bei Erwerbslosen zumindest zu übernehmen, dass Freiwilligkeit und Zustimmungsbereitschaft der Menschen eine der Grundvoraussetzungen von gesundheitsförderlichen Interventionen zu sein haben. Ansonsten wären zusätzliche Stigmatisierung, Viktimisierung und Sanktionierung die möglichen Folgen.

7 vergl. Bellwinkel, Kirschner 2011

8 Wie nun schon nahezu ›traditionell‹, gilt die Forderung nach Zielgruppenspezifität und -ausrichtung allgemein für die Prävention und Gesundheitsförderung, welcher allerdings nach wie vor häufig eine gegenteilige Praxis entgegensteht, wenn – wie beim Gros insbesondere kassengetragener Gesundheitsförderung – Angebote zielgruppenunspezifisch offeriert und dann mit entsprechendem Mittelschichtsbias »nachfrageseitig« in Anspruch genommen werden. Eine solche Ausrichtung ginge und geht umso mehr bei Erwerbslosen an deren Lebenswirklichkeit und ›Kontext‹ vorbei.

9 vergl. Bellwinkel, Kirschner 2011

Ohnehin besteht bei der Konzeptionierung und Durchführung von gesundheitsbezogenen Interventionen für Erwerbslose die Herausforderung und Schwierigkeit, dass Interventionen sensibel darauf auszurichten sind, keine Barrieren durch die Gefahr entstehen oder sich verstärken zu lassen, dass sich Erwerbslose zusätzlich zu ihrer sozialen Rolle nun auch als ›krank‹ stigmatisiert fühlen, wie dies durchaus eine Praxiserfahrung aus gesundheitlichen Interventionsprojekten bei Erwerbslosen ist.

3.2 Braucht es zweier verschiedener Kanonenkugeln in der Evaluation von Gesundheitsförderung für und bei Erwerbslosen?

Kirschner und Elkeles (2006) stellten bereits fest: »Was wir in der Gesundheitsförderung bei Arbeitslosen dringend brauchen, sind Maßnahmen und Projekte, die in der Konzeption und Implementation sowie Durchführung dem "state of the art" der Interventionsforschung entsprechen und evaluativ sauber auf ihre Wirksamkeit getestet werden« (ebda., S. 109).

Was das bedeuten *muss*? Oben war bereits gesagt worden, dass hier nicht der Anspruch besteht, hierzu etwa ein Kochbuch der Evaluation zu unterbreiten, dessen Anwendbarkeit von der Art und Anzahl der verfügbaren Kochtöpfe, -geräte und -utensilien abhängig sei, vielmehr sind die *Ingredienzien* oben in diesem Beitrag benannt worden: dieser In gredienzien hat sich der Koch je nach jeweiliger Mahlzeit in einer passenden Küche in entsprechend von ihm hergestellter Mixtur zu bedienen, damit ein adäquates Ergebnis erzielbar wird. Das ist es also, was es bedeuten *kann*. Anders gesagt: Evaluationen sind stets maßgeschneidert zu entwerfen – was dabei zu berücksichtigen ist, hängt vom jeweiligen Gegenstand und jeweiligen Kontext ab.

Wir können bzw. wollen uns daher abschließend darauf beschränken, einige Grundsätze zu benennen. Dabei erweitern wir die Perspektive von derjenigen des Kochs zu derjenigen des Restauranttesters.

Ein typischer Fehler wäre es, mehr Kanonenkugeln ins Spiel zu bringen, als zur Beobachtung und Beurteilung der sozialen Kausalität benötigt werden. Nicht nur, dass zuviel Kanonenmaterial, sprich die Hinzuziehung von Kontrollgruppen, den Evaluations-Einkaufspreis beim Auftraggeber unnötig erhöhen würde(n). Sondern die Beobachtung und Beurteilung des Mechanismus-Kontext-Wandel-Prozesses im Rahmen eines Programms könnte auch beeinflusst und gestört werden, wenn die soziale Realität durch eine fixe Versuchsanordnung interferiert wird. Eine – womöglich noch randomisierte – Zuteilung von Subjekten zu Interventionsobjekten (Treatment) einerseits und Placebos (No Treatment) stellt einen Eingriff in die soziale Realität dar, deren Konsequenzen unabsehbar sind. Im noch günstigsten Falle könnte die Kontrollgruppe ihren Ausschluss aus dem Programm als einen Ausschluss von der vielversprechenden Mahlzeit empfinden, was Einflüsse auf die in Interviews gegebenen Antworten haben dürfte. Solche Interferenzen zwischen Untersuchungs- und Kontrollgruppe dürften zudem je nach Kontext und setting mal schwächer, mal stärker ausfallen: Im Kontext der Job-Center mit den dort stets drohenden Sanktionsmöglichkeiten dürften solche Interferenzen besonders stark sein.

Da irgendein Vergleich (Beobachtung des Wandels), und zwar zwischen ›Soll‹ und ›Ist‹, bei der Evaluation jedoch stets vorzunehmen sein wird, wird der Koch nicht umhin kommen, dem Restauranttester zu ermöglichen, den Zustand vor und nach der Applikation der Mahlzeit zu beobachten und zu bewerten. Außerhalb der Gastronomie nennt man diese Vorgehensweise das Design eines Vorher-Nachher-Vergleichs. Koch bzw. Restauranttester werden wissen, dass hier die Wahl der Messzeitpunkte wichtig sein wird, um den vermuteten Mechanismus verifizieren oder falsifizieren zu können. Auch wird zumindest der Restauranttester bereits aus der Zeit der Tätigkeit seines Vorgängers bei den Hawthorne-Werken zur Zeit von Roethlisberger et al. (1966) wie jeder Sozialforscher wissen und entsprechend berücksichtigen, dass allein die Kenntnis, an einem Experiment und einer Beobachtung teilzunehmen, die Wahrnehmung beeinflusst. Human factors zu kennen, hält ihn jedoch weder vom Kochen noch vom Restauranttest ab, sondern sensibilisiert seine Aufmerksamkeit und Geschicklichkeit, die Programmwirkung in deren Kontext richtig kausal einschätzen zu können.

Last but not least wird es der Koch bzw. Restauranttester vermeiden, ausschließlich nach Ergebnissen wie dem Verdauungsgrad der Speisen pro

Zeiteinheit oder dem Sättigungsgrad in Abhängigkeit von der applizierten Kalorienzahl systematisch zu fragen (wenngleich es immerhin ein deutlicher Fortschritt ist, dass die heutige Evaluation der Gesundheitsförderung bei Erwerbslosen tatsächlich nunmehr auch und sogar vorwiegend outcomes misst). Als guter Koch, Restauranttester, Evaluator wird er wissen und anwenden, dass mindestens ebenso auch danach zu fragen und kausal zu interpretieren ist, unter welchen Bedingungen und wie die Ergebnisse zustande gekommen sind. In dieser sozialen Realität nämlich, in diesem Kontext, findet der soziale Wandel durch das Programm statt, falls er denn stattfindet.

Erst recht wird dies wichtig, wenn mehrere Köche in u. U. Subunternehmen beteiligt waren. Denn man muss nicht unbedingt Evaluator sein, um zu wissen, dass zu viele Köche den Brei verderben können. Werden etwa durch Subköche Sättigungs- und Zufriedenheitsziffern übermittelt, kann der Koch bzw. Restauranttester die soziale Kausation des durch das Beköstigungsprogramm erzielten Wandels nicht ohne gute Kenntnisse von Prozess und Kontext beurteilen. Quod est demonstrandum.

Literatur

Bhaskar R (1975) A Realist Theory of Science. Harvester, Brighton

Elkeles T (2001) Arbeitslosigkeit und Gesundheitszustand. In: Mielck A, Bloomfield K (Hrsg) Sozialepidemiologie. Eine Einführung in die Grundlagen, Ergebnisse und Umsetzungsmöglichkeiten. Juventa, Weinheim München, S 71–82

Elkeles T (2003) Arbeitende und Arbeitslose. In: Schwartz F, Badura B, Busse R et al. (Hrsg) Das Public Health-Buch. Gesundheit und Gesundheitswesen. 2. Auflage. Urban & Fischer, München Jena, S 653–659

Elkeles T, Beck D (2010) Evaluation von Betrieblicher Gesundheitsförderung – mehr als ein Datenvergleich. In: Faller G (Hrsg) Lehrbuch Betriebliche Gesundheitsförderung. Verlag Hans Huber, Bern, S 156–164

Elkeles T, Bormann C (1999) Arbeitslose. In: Bundesvereinigung Gesundheit e. V. (Hrsg) Gesundheit: Strukturen und Handlungsfelder. Luchterhand Verlag, Neuwied, II 6, S 1–20

Elkeles T, Bormann C (2002) Arbeitslose. In: Homfeldt HG, Laaser U, Prümel-Philippsen U et al. (Hrsg) Gesundheit: Soziale Differenz – Strategien Wissenschaftliche Disziplinen. Luchterhand, Neuwied Krifel, S 11–28

Elkeles T, Kirschner W (2004) Arbeitslosigkeit und Gesundheit. Intervention durch Gesundheitsförderung und Gesundheitsmanagement – Befunde und Strategien. Reihe Gesundheitsförderung und Selbsthilfe, Bd. 3 (Hrsg: BKK Bundesverband). Wirtschaftsverlag NW, Bremerhaven

Elkeles T, Kirschner W (2005) Unemployment as a determinant of health. In: Georgieva L, Burazeri G (Eds) Health determinants in the scope of new public health. Hans Jacobs Publishing Company, Sofia Lage, S 96–131

Elkeles T, Michel-Schwartze B (2009) Gesundheitsförderung in der Fortbildung für Fallmanager und Arbeitsvermittler. In: Hollederer A (Hrsg) Gesundheit von Arbeitslosen fördern! Ein Handbuch für Wissenschaft und Praxis. Fachhochschulverlag, Frankfurt am Main, S 230–260

Elkeles T, Seifert W (1992) Arbeitslose und ihre Gesundheit – Langzeitanalysen mit dem Sozio-ökonomischen Panel. Soziale Welt 43 (3): 278–300

Elkeles T, Seifert W (1993a) Arbeitslose und ihre Gesundheit – Langzeitanalysen für die Bundesrepublik Deutschland. Soz Praventivmed 38 (2): 148–155

Elkeles T, Seifert W (1993b) Unemployment and Health-Impairments: Longitudinal Analyses from the Federal Republic of Germany. European Journal of Public Health 3 (1): 28–37

Elkeles T, Seifert W (1993c) Migration und Gesundheit. Arbeitslosigkeits- und Gesundheitsrisiken ausländischer Arbeitsmigranten in der Bundesrepublik Deutschland. Sozialer Fortschritt 42 (10): 235-241

Elkeles T, Seifert W (1996) Immigrants and Health. Unemployment- and Health – Risks of Labour Migrants in the Federal Republic of Germany, 1984–1992. Soc Sci Med 43 (7): 1035–1047

Goffman E (1975) Stigma. Über Techniken der Bewältigung beschädigter Identität. Suhrkamp, Frankfurt am Main

Harré R (1972) The Philosophies of Science. Oxford University Press, Oxford

Hellstern GM, Wollmann H (Hrsg) (1984) Handbuch zur Evaluierungsforschung. Bd. 1. Westdeutscher Verlag, Opladen

Hume D (1989) Ein Traktat über die menschliche Natur. Bd. 1 (2. Auflage von 1904). Meiner, Hamburg

Jahoda M, Lazarsfeld PE, Zeisel H (1973) Die Arbeitslosen von Marienthal. Suhrkamp, Frankfurt

Kieselbach T, Beelmann G (2006) Arbeitslosigkeit und Gesundheit: Stand der Forschung. In: Hollederer A, Brand H (Hrsg) Arbeitslosigkeit, Gesundheit und Krankheit. Verlag Hans Huber, Bern u. a., S 13–31

Kirschner W, Elkeles T (2006) Eine aktuelle Bestandsaufnahme von deutschen Projekten zur Gesundheitsförderung von Arbeitslosen – Probleme, Forschungs- und Entwicklungsbedarfe. In: Hollederer A, Brand H (Hrsg) Arbeitslosigkeit, Gesundheit und Krankheit. Verlag Hans Huber, Bern u. a., S 97–112

Kratzer N (2003) Arbeitskraft in Entgrenzung. Grenzenlose Anforderungen, erweiterte Spielräume, begrenzte Ressourcen, edition sigma, Berlin

Kromrey H (2001) Evaluation – ein vielschichtiges Konzept. Begriff und Methodik von Evaluierung und Evaluationsforschung. Empfehlungen für die Praxis. Sozialwissenschaften und Berufspraxis 24 (2): 105–132

Mill JS (1961) A System of Logic. Longman, London

Mohr G (2010) Erwerbslosigkeit. In: Kleinbeck U, Schmidt KH (Hrsg) Enzyklopädie der Psychologie: Themenbereich D, Serie 3, Band 1 Arbeitspsychologie, 2. Auflage. Hogrefe, Göttingen, S 471–519

Paul KI, Hassel A, Moser K (2006) Die Auswirkungen von Arbeitslosigkeit auf die psychische Gesundheit: Befunde einer quantitativen Forschungsintegration. In: Hollederer A, Brand H (Hrsg) Arbeitslosigkeit, Gesundheit und Krankheit.Verlag Hans Huber, Bern u. a., S 35–51

Pawson R, Tilley N (1997) Realistic Evaluation, Reprint 2009, Sage, London

Roethlisberger FJ, Dickson WJ, Wright HA (1966) Management and the Worker, An Account of a Research Program. Conducted by the Western Electric Company, Hawthorne Works, Chicago [1939], 14. Auflage. Harvard University Press, Cambridge

Rossi PH, Freeman HE, Hofmann G (1988) Programm-Evaluation. Einführung in die Methoden angewandter Sozialforschung. Enke, Stuttgart

Seligman MEP (1979) Erlernte Hilflosigkeit. Urban und Schwarzenberg, München Wien Baltimore

Teil 2
Praxiseinsichten:
Evaluation komplexer Interventionen als reflexiver Entwicklungsprozess

Die Evaluation von Setting-Interventionen mit dem Instrumentarium der Krankenkassen: Erfahrungen mit einem System zur Projekt-, Dach- und Metaevaluation

Thomas Kliche

1 Komplexe Interventionen der Krankenkassen

Die Krankenkassen sind mit Abstand die wichtigsten Träger komplexer Interventionen zur Gesundheitsförderung in der Bundesrepublik Deutschland. Sie gaben 2010 knapp 23 Millionen Euro für Projekte des Setting-Ansatzes aus, die rund 2,4 Mio. Menschen in etwa 30.000 Lebenswelten bzw. Organisationen direkt erreichten, darunter 9.000 allgemeinbildende Schulen und 16.000 Kitas (Schempp et al. 2012). Die Krankenkassen tragen nicht allein quantitativ die meisten komplexen Interventionen, sondern sie haben auch qualitativ ein Interventions- und Evaluationssystem aufgebaut, das effizient die Bevölkerungsgesundheit verbessern soll (GKV-Spitzenverband 2010; Stuppardt, Wanek 2009).

2 Projekt-, Dach- und Metaevaluation im Evaluationssystem der Krankenkassen

Eine Schlüsselstellung in diesem System nehmen die Evaluationsinstrumente ein. Sie sollen a) die Einzelprojekte zeitnah über Erfolge orientieren und b) die Erfahrungen bundesweit aggregieren, um die Krankenkassen über erfolgreiche und effiziente Vorgehensweisen zu informieren und dadurch die Interventionsqualität kontinuierlich zu verbessern. Dafür sind nicht regelmäßige jährliche Datenzusammenführungen zur Gesamtwirksamkeit der vielen heterogenen Projekte entscheidend, also Dachevaluationen im engeren Sinne. Erforderlich sind darüber hinaus systematische Metaevaluationen. Diese beruhen auf Stichproben oder Teilgesamtheiten aller verfügbaren Projekte und Befragten, die unter theoretischen oder methodischen Fragestellungen ausgewählt und verglichen werden. Metaevaluationen können Auskunft geben über die Wirksamkeit von:

- Vorgehensweisen (Zusammenstellung, Vollständigkeit und Abfolge der Interventionselemente, z. B. Schwerpunkt auf Maßnahmen zur Vernetzung der Schulen oder zur Personal- und Organisationsentwicklung oder für spezifische Gesundheitsziele der Schüler/-innen, etwa Soziale Unterstützung oder diverse Risikoverhalten);
- Dosis-Effekten (z. B. Projektdauer, Exposition der einzelnen Schüler/-innen für bestimmte Maßnahmen, Breite der Maßnahmen in bestimmten Klassen-, Alters- oder Genderstufen);
- Gestaltung der Teilschritte (z. B. Einsatz evidenzgestützter oder selbstgebastelter Programme, Qualitätssicherung des Vorgehens, Start der spezifischen Maßnahmen zur Gesundheitsförderung erst nach Errichtung eines vollständigen Projekt- und Partizipationsapparats an den Schulen oder rasch nach Projektbeginn zwecks Erreichung greifbarer, motivierender Erfolge);
- konfundierenden Ausgangsbedingungen, d. h. Mediatoren- und Moderatoreneffekte (z. B. Einrichtungsart, -größe, Organisationskultur, Teilnehmermerkmale).

3 Anforderungen an das Evaluationssystem der Krankenkassen

Um für beide Ebenen – die der Einzelprojekte und die der Dach- bzw. Metaevaluation – geeignet zu sein, muss das Evaluationssystem der Krankenkassen einfache, aber hohe Anforderungen erfüllen (Kliche et al. 2011b).

1. Effekte vergleichsfähig beschreiben, den Blick also auf generisch relevante Outcomes richten.
2. direkt und indirekt relevante Dimensionen einbeziehen, also sowohl Gesundheits-Outcomes als auch Outputs (z. B. Zahl von Informations-

veranstaltungen, Veränderungen der Organisationsstruktur);
3. Struktur-, Prozess- und Ergebnisqualität in den Blick nehmen, um Beziehungen zwischen Ausgangslagen, Vorgehen und Outcomes herzustellen;
4. für alle Zielgruppen verwendbare Instrumente bieten, also ohne Bildungs-, Alters- oder Gender-Einschränkungen,
5. Schlüsselindikatoren berichten, d. h. Kennziffern hoher externer Validität mit Ausstrahlung für viele gesundheitlich bedeutsame Verhalten und -dimensionen;
6. parallele Instrumente für verschiedene Vorgehensweisen bieten, damit die Ergebnisse mit den Outcomes anderer Felder – der Betrieblichen Gesundheitsförderung und des Individualansatzes – verglichen werden können;
7. Instrumente hoher Brauchbarkeit zusammenstellen. Hierzu sind – wie bei der Qualitätsentwicklung von Prävention und Gesundheitsförderung generell – eine Reihe von Gütekriterien zu beachten (Kliche et al. 2009):
   a) Sparsamkeit bei Einsatz und Auswertung,
   b) Geschwindigkeit, d. h. wenig Zeit für Einarbeitung, Datenerhebung und Auswertung, so dass rasche Rückmeldungen möglich sind,
   c) einfache Handhabung, d. h. vor allem leichte Verständlichkeit der Instrumente und ihrer Gebrauchsanweisungen sowie der Rückmeldungen und Befunde,
   d) Nutzbarkeit für mehrere Organisationszwecke, insbesondere rasche Beurteilung aktueller Teilvorhaben, Beurteilung der Wirksamkeit von Teilschritten, Qualitätsverbesserung des Projekts, Dokumentation von Erfolgen für die Zielgruppen,
   e) sowie selbstverständlich die Hauptgütekriterien wissenschaftlich fundierter Aussagen – Unparteilichkeit, Zuverlässigkeit und Gültigkeit, und als deren Teilaspekt namentlich Normierbarkeit (d. h. die Verankerung der Befunde an Vergleichswerten der Bevölkerung oder relevanter Teilgruppen, um ihre Bedeutung zu bewerten).

4 Die Entwicklung des Evaluationssystems

Um diesen Anforderungen gerecht zu werden, wurden zunächst die besonderen Vorgehensweisen des Individualansatzes und der Betrieblichen Gesundheitsförderung – obgleich sie häufig als integraler Bestandteil anderer komplexer Interventionen zu finden sind – in eigenen Teilinstrumentarien erfasst (Kliche et al. 2010b; Kliche et al. 2011c). Die Wirkungsabschätzung komplexer Interventionen wurde am Beispiel Schule/Kita in einem Pilotprojekt der Krankenkassen in Niedersachsen, Sachsen-Anhalt und Rheinland-Pfalz 2003–2007 entwickelt und erprobt (Heinrich et al. 2006; Kliche et al. 2010a). Instrumententestung und -optimierung erfolgten also in einem realen Interventionsfall.

Die Koordination und überwiegend auch die Durchführung der Schulprojekte lagen bei den Landesvereinigungen für Gesundheit, die Fachkräfte mit den erforderlichen Qualifikationen in Gesundheits- und Organisationsentwicklung bereitstellten. Sie wurden von Landes- und Regionalexperten und Fachkräften der Krankenkassen bei Beratung, Vernetzung und Akquise von Zusatzmitteln für die beteiligten Schulen unterstützt. Das Projekt wurde nach 2007 schul- und länderspezifisch fortgesetzt, in Niedersachsen beispielsweise durch Einführung der Balanced Score Card in einer spezifisch für Schulen ausgelegten Fassung (Liersch et al. 2012).

Am Pilotprojekt beteiligt waren 55 Schulen und 2 Kitas. In den Einrichtungen einbezogen waren die Kollegien, Schüler- und Elternvertretungen, Schulleitungen und Gesundheitsmoderator/-innen sowie – soweit erforderlich – weitere externe Experten/Expertinnen (z. B. aus Lehrerbildung, Prävention und Gesundheitsförderung, Drogenberatung).

5 Setting-Ansatz, Interventionskerne und länderspezifische Strategien

Komplexe Interventionen deutscher Krankenkassen fußen auf dem Setting-Ansatz der Weltgesundheitsorganisation (GKV-Spitzenverband 2010). Dieser zielt bekanntlich auf die gesundheitsgerechte Umgestaltung von Lebenswelten ab, sucht mehrere Zielgruppen einzubeziehen, kombiniert Verhaltens- mit Verhältnisprävention, folgt den Stra-

tegien der Partizipation, des Empowerment durch Vermittlung von Problemlösungskompetenzen (life skills) und der Ressourcenstärkung (z. B. der Erhöhung von Selbstwirksamkeitserwartung) und arbeitet mit einem breiten, feldabhängigen Repertoire an Einzelinterventionen (»Werkzeugkasten«) (Dooris 2009; Engelmann, Halkow 2008).

Als Interventionskern ergab sich daraus für die untersuchten Schulprojekte: Eine Schulkonferenz beschloss gesundheitsbezogene Zielsetzungen und Schwerpunkte für das Schulprogramm. Sodann wurde in der Schule eine handlungsfähige Projekteinheit errichtet (Steuerkreis oder Gesundheitszirkel, aber auch Multiplikatorengruppen von so genannten »Gesundheitsmoderator/-innen«). Eine Eingangserhebung ermittelte Gesundheitsstatus und Problem, Wünsche und Erwartungen an die Einrichtung und das Projekt. Auf den festgestellten Präferenzen der Zielgruppen bauten Teilschritte auf, die lokal ausgewählt und in unterschiedlicher Weise umgesetzt wurden, darunter Umbauten (Sportanlagen, Trinksäulen, Pausenhof), Anschaffungen (Klettergurte für Bäume, Spiel- und Sportkisten), Veranstaltungen (Gesundheits-/Projektwoche/-tage, Informationsabende für Eltern) und gesundheitsbezogene Themen im Unterricht (z. B. zu Suchtproblemen). Über Art, Qualität und Dosis der Maßnahmen entschieden die Steuerkreise der Einrichtungen.

Im Rahmen der Kerninterventionen realisierten die Bundesländer eigene Strategien: Errichtung eines Netzwerks gegenseitiger Unterstützung und Multiplikatoren-Qualifikation, Übertragung bewährter Werkzeuge des Gesundheitsmanagements auf Bildungseinrichtungen sowie Umsetzung Betrieblicher Gesundheitsförderung in Schulen.

6 Der Forschungsplan für Systementwicklung und -einsatz

Da die Systementwicklung zugleich der Erprobung der Instrumente diente, umfasste der Forschungsplan im Kern den Plan eines realen Systemeinsatzes: eine Beobachtungsstudie mit Messpunkten zu Projektbeginn und -ende nach ungefähr zwei Jahren.

Hinzu traten methodisch aufwendigere Schritte zur Absicherung der wissenschaftlichen Güte des Evaluationssystems. Hierzu zählten Dosisabschätzungen, psychometrische Analysen und qualitative Begleitforschung (81 Interviews zu Prozessqualität, Effekten und Wirkmechanismen des Projekts) zur Dokumentation von Kausalketten, Aussagekraft der Instrumente und externer Validität. Ein dritter Messpunkt wurde im dritten Projektjahr eingerichtet, um die Nachhaltigkeit gesundheitsbezogener Veränderungen in den Schulstrukturen zu dokumentieren. Die Ergebnisse wurden nicht allein in Rohwertform berechnet, sondern auch verankert; dazu wurde die Referenzwertdifferenz zur individuellen Alters- und Geschlechtsgruppe der Befragten errechnet, so dass prozentuale Abweichungen von der Gesamtbevölkerung als Kontrollgröße beobachteter Veränderungen herangezogen werden konnten. Schließlich wurden Ausfall- und Verzerrungsanalysen berechnet, um den Total Survey Error abzuschätzen und zu minimieren.

Mit diesen Operationen konnte belegt werden, dass das Evaluationssystem der Krankenkassen die wichtigsten Vergleiche prinzipiell ermöglicht, die zur Beurteilung komplexer Interventionen erforderlich sind, nämlich zwischen

- Messpunkten (über die Zeit),
- Zentren bzw. Einrichtungen (über Ausgangslagen),
- soziodemografischen Teilnehmergruppen (über Alter, Gender, soziales Umfeld),
- Vorgehensweisen (über Dosis, Interventionsdichte und -dauer), und
- Risikogruppen (über Zielgruppen im Vergleich zur jeweiligen Alters- und Gendergruppe in der Gesamtbevölkerung).

7 Erfasste Parameter

Die zentralen Variablen des Evaluationssystems geben die folgenden Tabellen wieder. Sie sind auf mehrere zielgruppengerechte Instrumente von wenigen Seiten Länge verteilt. Die Interventionsdokumentation ist für jedes abgrenzbare Teilvorhaben gefordert; Strukturerhebungen sind zu Projektbeginn und -ende zwei Jahre später sowie ein Jahr darauf sinnvoll, die Teilnehmerbefragung mindestens zu Beginn und Ende eines Vorhabens.

Tabelle 1

Interventionsdokumentation (Antworten je Einzelintervention zum Ankreuzen oder Ausfüllen, Mehrfachangaben möglich)
Zielebene der Veränderungen: Vorlauf, Initiierung, Projektmotivation; Entscheidungsstrukturen; räumliche Gestaltung; Abläufe im Schulalltag; Unterrichtspraxis; Kommunikation der Schüler/-innen; Kommunikation Schüler/-innen-Lehrkräfte; Kommunikation unter Lehrkräften; Kommunikation Schulleitung-Lehrkräfte; Kooperation Schule-Eltern; Kooperation mit schulexternen Partnern; Versorgungsangebote (z. B. Beratung, Ernährung); individuelle Kenntnisse und Fähigkeiten; Sonstiges
Themen: Bewegung; Stress/Entspannung; Ernährung; Genuss-/Suchtmittel; Arbeitsplatz Schule und Gesundheit; Umwelt und Gesundheit; soziale Beziehungen und Gesundheit; Führung/Moderation; Sonstiges
Interventionsart: Analyse (Begehung, Befragung), Gesundheitszirkel, Gesundheitstage/-woche; Fort-/Weiterbildung; Arbeits-/Projektgruppe; Präventionskurs; Informationsveranstaltung; individuelle Beratung; Sonstiges
Arbeitsaufwand: Zahl der Termine/Treffen; deren mittlere Dauer in Stunden; Zahl regelmäßig beteiligter Personen; geschätzte Arbeitszeit pro Person; Anzahl Beteiligter aus externer Institution
erreichte Personen: Schüler/-innen; Lehrkräfte; nichtpädagogisches Personal; Schulleitungen; Eltern; externe Kooperationspartner

Tabelle 2

Strukturerhebung (Dimensionen aus je mehreren Items, Ordinalskalierung (0 = nicht begonnen; 1 = geplant/in Vorbereitung; 2 = teilweise erreicht; 3 = voll erreicht)
Verankerung von Gesundheitsförderung in der Einrichtung
Gesundheits- und empowermentbezogene Fortbildung
Planungsqualität
Steuerung von Gesundheitsförderung
Partizipation aller Gruppen an Planung und Gestaltung von Gesundheitsförderung
Vielfalt der gesundheitsbezogenen Angebote und Maßnahmen
(kommunale) Vernetzung
Qualitätssicherung der Gesundheitsförderung
Soziale Verantwortung der Einrichtung
Ausstattung der Einrichtung insgesamt und unter Gesundheitsaspekten

Tabelle 3

Gesundheitsparameter der Zielgruppen (unterschiedliche Item-Zahl und Skalierung)
Subjektive Gesundheit und Gesundheitliche Lebensqualität
Arbeits- und Organisationszufriedenheit
Bewertung der Ausstattung der Einrichtung
Unterstützung durch Lehrer/Vorgesetzte
Einzelbeschwerden
Einzelbelastungen
Schulische Belastungen
Körpergröße, Gewicht (BMI)
Krankheitstage
Gesundheitsverhalten: Bewegung, Ernährung, Stressbewältigung, Rauchen, Umgang mit Alkohol

8 Ausgewählte Erfahrungen zur Evaluation komplexer Interventionen

8.1 Einfache, sachgerechte Instrumente können Projektverläufe und Wirkungsebenen differenziert abbilden

Die Systemerprobung dokumentierte, dass die Instrumente und Auswertungen des Evaluationssystems der Krankenkassen informative Ergebnisse über komplexe Interventionen in Settings erbringen (Kliche 2008; Kliche et al. 2010a). Einige Ergebnisse waren für die Einzel- und Dachevaluation komplexer Interventionen generell von Bedeutung:

8.2 Heterogene Wirksamkeit und starke Zentren-Unterschiede

Der Ansatz brachte Effekte unterschiedlicher Größe und Richtung auf die Ausgangslage. Veränderungen der drei wichtigsten Outcome-Dimensionen – Strukturverbesserungen, Lehrer- und Schülergesundheit – wiesen auch nicht an allen Schulen in die gleiche Richtung. Das ist plausibel, wie die qualitative Begleitforschung zeigte: Interventionen kosten Zeit und Energie, und wenn eine Verbesserung der Schülergesundheit oder der Schulstrukturen Vorrang hat und die Lehrkräfte mit Mehrarbeit und Auseinandersetzungen belastet, so können sich Arbeitsklima und Gesundheit verschlechtern. Unterschiede entstanden somit durch das Vorgehen vor Ort und auf Länderebene.

Als weitere statistisch eigenständige Quellen unterschiedlicher Verläufe erwiesen sich u. a. Einrichtungs- und Kollegiengröße sowie die Einrichtungsart: Weiterführende Schulen profitierten mehr vom Projekt. Effekte von Einrichtungsmerkmalen und Organisationskultur sind auch international aus Schulstudien bekannt (Griebler et al. 2009; Richard et al. 2012). Diese Befunde haben weit reichende methodische Konsequenzen (s. u.).

8.3 Dosiseffekte und Schwächen der Interventionserfassung

Das Vorgehen an den Schulen wurde mit einem Kurzbogen erfasst (Tab. 1); er ermöglichte eine Dosis-Abschätzung und erbrachte Kausalitätshinweise: Schulen, die die höchsten Effekte erzielten, setzten etwa doppelt so viel Zeit für das Projekt ein wie andere.

Die Nachinterviews offenbarten jedoch, dass der Kurzbogen nur lückenhaft eingesetzt und nicht gleichförmig ausgefüllt wurde. Einerseits befürchteten die Ausfüllenden eine Kontrolle ihrer Arbeit und Produktivität, insbesondere in Fällen ›wilder‹, selektiver Spontanadaption des Programms oder seiner Bausteine. Andererseits fiel ihnen aufgrund einer latenten Entwertung der eigenen Arbeit die Abgrenzung zwischen Projektaktivitäten und »selbstverständlichen« Alltagsaufgaben schwer: Das – vermeintlich selbstverständliche, freiwillige – Engagement vieler Eltern, Schüler/-innen und Lehrkräfte in deren Freizeit wurde unregelmäßig in die Daten einbezogen. Schließlich spiegelten die Evaluationsdaten auch eine Entwertung informeller Arbeit. Wo die Projektleitung viel Arbeit übernommen hatte, die anfangs nicht eingeplant war, schämte sie sich mitunter, dies anzugeben, weil sie das Kollegium nicht zu motivieren vermocht hatte und damit ihre Kompetenzen und ihr Status in Zweifel standen. Insgesamt unterschätzten die Bögen daher klar die im Projekt aufgewendete Arbeitszeit.

8.4 Unterschiedliche Vorgehensweisen und Verläufe

Die gemeldeten Aktivitäten (Einzelinterventionen) der Projekte offenbarten deutliche Unterschiede der Verläufe in den drei Bundesländern, nicht allein zeitlich versetzt, sondern auch mit unterschiedlichen Schwerpunkten der Aktivitäten zu Beginn und im Verlauf. Auch wo Kerninterventionen definiert sind, können in komplexen Programmen verschiedene »Teilfassungen« stecken, die gesondert auf Effekte untersucht werden müssen.

8.5 Proximale und distale Effekte

Die Effektgrößen der erzielten Veränderungen zeigten ein Muster: Organisationsmerkmale (z. B. Schulprogramme) lassen sich vergleichsweise unaufwendig und daher rasch und in vielen Einrichtungen ändern. Kleinere Zielgruppen

mit enger Bindung an die Einrichtung lassen sich leichter erreichen als größere, die weniger Zeit im Setting zubringen (Lehrkräfte und Schüler/-innen bzw. deren Familien). Bei komplexen Interventionen ist somit von proximalen, raschen und leichter erzielbaren Effekten einerseits, von distalen Effekten andererseits auszugehen (Kliche et al. 2006).

9 Folgerungen und Empfehlungen

9.1 Stringente Prozessbeschreibung

Eine gesundheitsökonomische Kosten-Nutzen-Abwägung komplexer Projekte ist nur anhand genauer, einheitlicher Prozessdaten möglich. Bereits der einfache Kurzbogen über Einzelinterventionen brachte wichtige Aufschlüsse und kann ökonomisch eingesetzt werden. Die Beteiligung an der Evaluation war für die Projekte obligatorisch, doch weder diese Verpflichtung noch Instruktionen noch ein einfaches Instrument genügten zur Erhebung genauer Prozessdaten. Evaluationen komplexer Interventionen sollten daher mit diesen und weiteren Mitteln die Datenqualität absichern und benötigen hierfür starke Unterstützung der Projektleitungen und -träger sowie Ressourcen.

9.2 Zentreneffekte: 'Confounder' und erforderliche Alternativen zu RCTs

Einrichtungsgröße und Organisationskultur, auch das soziale Kapital in einer Einrichtung und deren Umfeld erwiesen sich als signifikante Ergebnisdeterminanten. Dies hat weit reichende Folgen. Da die Effekte auf Cluster-Ebene entstehen, ergeben sich für randomisierte Kontrollstudien (RCT) aus methodischen und ökonomischen Gründen schwer lösbare Probleme, u. a. folgende:

a) Die Randomisierung einzelner Teilnehmer ist unseriös, da ganze Einrichtungen (Zentren) von Unterschieden wie Interventionen betroffen sind.
b) Zufallsgenerierte Kontrollcluster von Zentren würden unbezahlbar große Stichproben zur Abbildung aller Confounder und Interaktionen auf Einrichtungsebene erfordern, z. B. eine Zufallsstichprobe von Hauptschulen eines Bundeslandes im Vergleich zu Interventionsschulen mit hinreichend große Konfidenzintervallen auf Einrichtungsebene.
c) Zufallsgenerierte Kontrollcluster wären aus mikropolitischen Gründen unmöglich; so würden sich Schulen mit hohem Problemdruck in einer Wartekontrollgruppe während einer zweijährigen Intervention rasch vorher andere Hilfe beschaffen (solche Zusatzhilfen wurden auch in den Projektschulen in Anspruch genommen).

Somit sind Alternativen zu den – für komplexe Interventionen naiven – RCT erforderlich. Nach den Erfahrungen mit dem System der Krankenkassen sind drei Wege denkbar:

1. gezieltes Matching der Interventionseinrichtungen mit Kontroll-Zentren, z. B. Schulen ähnlicher Art, Größe und Umfelder. Hierfür sind hinreichende Kenntnisse der wichtigsten Confounder erforderlich, um diese in Vergleichsgruppen-Designs gezielt zu variieren.
2. Verankerung und Angabe von Referenzwertdifferenzen: Sie wurden im System der Krankenkassen »eingebaut« und ermöglichen einen höchst ökonomischen Vergleich mit der Gesamtbevölkerung. Hierfür sind alters- und genderspezifische Referenzdaten als individuelle Vergleichswerte sowie eine komplexere Auswertung erforderlich.
3. Metaevaluation und Benchmarking: Das System der Krankenkassen ermöglicht Vergleiche von Einrichtungen mit unterschiedlichen Vorgehensweisen, Ausgangslagen und Organisationsmerkmalen. Bei hinreichenden Datensätzen sind wichtige Fragen über komplexe Interventionen auf diese Weise zu beantworten. Hierfür ist eine im Kern einheitliche Projektevaluation mit anschließender Zusammenführung aller einzelnen Datensätze in einem zentralen Pool erforderlich; aus ihm können dann für alle Fragestellungen geeignete Vergleichsstichproben von Einrichtungen und deren Mitgliedern gebildet werden, z. B. von Schulen, deren Kollegien und Schüler/-innen. Diese Werte können auch als Benchmarks für neue Projekte zur Verfügung gestellt werden.

9.3 Erhebungszeitraum

Zur Beobachtung proximaler und distaler interagierender Effekte in komplexen Interventionen kann der Evaluationshorizont für die jeweils erwarteten Veränderungen evidenzgestützt begründet werden. Je nach Zielsetzung können sich einige Monate, aber auch mehrere Jahre ergeben.

9.4 Konkurrenz komplexer mit spezifischen Interventionen

Die erzielten Effektgrößen waren im Mittel sehr niedrig, in einzelnen Einrichtungen negativ. Damit konkurrieren Setting-Projekte bei klar im Vordergrund stehenden Problemlagen (z. B. Sucht, Nikotinprävention, kollegiales oder Klassenklima, Sexualaufklärung, Aggressivität) mit spezifischen Programmen. Diese vermögen in kürzerer Zeit höhere Effektgrößen erreichen, weil sie nicht eine ganze Lebenswelt umzustrukturieren suchen, sondern rasch und evidenzbasiert spezifische Ziele umsetzen (Lister-Sharp et al. 1999; Stewart-Brown 2006; Weare, Nind 2011). Die Entscheidung zwischen komplexen und spezifischen Interventionen ist auf zwei Ebenen zu treffen:

a) gesundheitsökonomisch: Hinreichend differenzierte Daten für eine Kosten-Nutzen-Abschätzung von Setting-Projekten, unter vollem Einbezug informeller Arbeit kann das System der Krankenkassen bei konsequenter Nutzung liefern.
b) pragmatisch für jedes Setting: In Situationen mit brüchiger Motivation der Zielgruppen kann einerseits Partizipation im Sinne des Setting-Ansatzes eine integrierende Strategie bieten, aber – wie die Daten zeigten – auch scheitern; andererseits können rasche greifbare Erfolge evidenzbasierter Programme den Einfluss der Gesundheitsförderung erhöhen.

Auch für solche Entscheidungen kann das System der Krankenkassen die Evidenzbasis schaffen.

9.5 Evaluationsstandards

Um die Potentiale von Dach- und Metaevaluationen voll zu nutzen, sind Standards erforderlich, die die Evaluation von Einzelprojekten einbeziehen. Sie sollten umfassen:

- zuverlässige, valide Erhebung von Art, Umfang und Qualität der Interventionen. Auf der Ebene der Teilschritte steht im System der Krankenkassen der Kurzbogen zur Interventionsbeschreibung zur Verfügung, auf der Ebene der Projektkonzeption z. B. »Qualität in der Prävention« (Töppich, Lehmann 2009). Einfache Projektentwicklungs-Tools oder die Honecker-artige Selbstbewertung selbstgesetzter Ziele (“goal attainment scaling”) sind hingegen unvalide.
- Einsatz normierter, validierter Indikatoren und Messinstrumente für Gesundheits- und Organisationseffekte. Diese Vorgabe würde die Praxis der Krankenkassen verändern. Deren Setting-Projekte werden zwar zu 74 % evaluiert, doch überwiegend mit Akzeptanzerhebungen unter den Projektbeteiligten, also Marktforschung, während Gesundheitsindikatoren nur bei 28 % der evaluierten Vorhaben in Betracht gezogen wurden (Schempp et al. 2012).
- Multivariate Modellierung: Simple Mittelwertvergleiche sind in komplexen Interventionen irreführend, wo Ausgangslagen sich unterscheiden; die Bedeutung von Moderatoren und Mediatoren hat in anderen Feldern des Gesundheitswesens als Prinzip der »fairen Vergleiche« Anerkennung gefunden (Farin et al. 2004). Für Ausfall-, Verzerrungs- und Confounder-Kontrolle, Zentren- und Mehrebenen-Effekte, Interaktionen von Wirkketten und auch für Teilgruppenvergleiche sind komplexe Auswertungen angezeigt. So wurde das System der Krankenkassen in der Erprobung auf Eignung für multivariate Varianzanalysen, Logistische Regressionen, Mehrebenenanalysen und Strukturgleichungsmodelle getestet.
- Benchmarks für Einzelprojekte: Multivariate Auswertungen eignen sich wegen hohen Aufwands und anspruchsvoller Interpretation nicht für die Erfolgsrückmeldung von Einzelprojekten. Hierfür sollten Benchmarks herangezogen werden (z. B. Verankerungswerte oder Vergleichswerte anderer Einrichtungen). Eine

Spannung zwischen potentiell irreführenden Rohmittelwerten und echten »fairen Vergleichen« wird also zunächst fortbestehen. Die Aussagekraft und Interpretierbarkeit von Mittelwertvergleichen in Einzelprojekten werden indes mit zunehmender Kenntnis der relevanten Einrichtungsmerkmale durch Metaevaluation über die Jahre zunehmend anwachsen und jene Spannung schrumpfen lassen.

- Datendokumentation: Der Nutzen von Dach- und Metaevaluationen und Benchmark-Findung kann nur ausgeschöpft werden, wenn die Daten in einheitlicher Weise einem zentralen Datenpool zugeführt werden. Dies ist bislang nicht gelungen. Ursachen dürften in Kostengründen, aber auch in der damit erzwungenen Vereinheitlichung und Transparenz der Evaluationen und Interventionserfolge liegen.

9.6 Zusatzbedarf bei komplexen Evaluationen

Für eine Reihe bislang ungewöhnlicher Methodenschritte sind höhere Mittel erforderlich, als sie Auftraggebern vertraut sind: Prozessabbildung (Strategie und Interventionsqualität), Abschätzung des Interventionsumfangs (Reichweite, Expositionszeit u. ä.), Sicherung der Validität durch interne und externe Vergleichsgruppen, Clusterzentren-Auswertungen mit Matching oder Verankerung oder Benchmarks, Moderatoren- und Mediatorenkontrolle für faire Vergleiche. Erforderlich sind ggf. weiter Rekrutierungs- und Organisationsaufwendungen für das Matching von Zentren. Empfehlenswert ist auch qualitative Begleitforschung zur Validitätssicherung (Erhellung von Wirkungskausalketten) und zur Ermittlung von Umsetzungserfahrungen, Hürden und unerwarteten »Nebeneffekten«.

9.7 Methodenentwicklung

Die Wissenschaft hat Grundlagenaufgaben für komplexe Interventionen zu Ende zu führen:

- die Entwicklung empirisch gestützter Taxonomien von Einzelinterventionen (z. B. Setting-Ansatz und spezifische Einzelprogramme, Operationalisierung von »Komplexität«).
- die Erhellung von Zentreneffekten, um die Verzerrung durch Ausgangslagen abzuschätzen und Regeln für systematische Matching-Stichproben und angemessene Benchmarks zu finden. Hierfür sind Reviews ein erster Schritt (Kliche, Goßmann 2011a).
- die Verbesserung der Alternativen zu RCTs.
- die Errichtung von Plattformen zur Datensammlung und Sekundäranalyse in Dach- und Metaevaluationen unter Mitwirkung von Versorgungsträgern, vor allem der Krankenkassen.

Literatur

Dooris M (2009) Holistic and sustainable health improvement: the contribution of the settings-based approach to health promotion. Perspect Public Health 129 (1): 29–36

Engelmann F, Halkow A (2008) Der Setting-Ansatz in der Gesundheitsförderung. Genealogie, Konzeption, Praxis, Evidenzbasierung. SP I 2008-302. Wissenschaftszentrum Berlin, Berlin

Farin E, Glattacker M, Follert P et al. (2004) Einrichtungsvergleiche in der medizinischen Rehabilitation. Zeitschrift für ärztliche Fortbildung und Qualität im Gesundheitswesen 98 (8): 655–662

GKV-Spitzenverband (Hrsg) (2010) Leitfaden Prävention. Handlungsfelder und Kriterien des GKV-Spitzenverbandes zur Umsetzung von §§ 20 und 20a SGB V vom 21. Juni 2000 in der Fassung vom 27. August 2010. In Zusammenarbeit mit den Verbänden der Krankenkassen auf Bundesebene – AOK-Bundesverband, BKK Bundesverband, IKK e. V., Spitzenverband der landwirtschaftlichen Sozialversicherung, Knappschaft, Verband der Ersatzkassen. GKV-Spitzenverband, Berlin

Griebler R, Dür W, Kremser W (2009) Schulqualität, Schulerfolg und Gesundheit. Ergebnisse aus der österreichischen "Health Behaviour in School-Aged Children"-Studie. Österr Zeitschr Soziol 34 (2): 79–88

Heinrich S, Kolbe M, Schwabe U et al. (2006) gesund leben lernen. Lebensräume gestalten – gesundes Handeln ermöglichen. Prävention. Jahrbuch für Kritische Medizin (43): 40–54

Kliche T (2008) Gesund Leben Lernen – Ergebnisse der externen Evaluation. In: AG-SpiK (Hrsg) Arbeitsgemeinschaft der Spitzenverbände der Krankenkassen und Medizinischer Dienst der Spitzenverbände der Krankenkassen (MDS), unter Beteiligung des GKV-Spitzenverbandes. Präventionsbericht 2008. Dokumentation von Leistungen der gesetzlichen Krankenversicherung in der Primärprävention und betrieblichen Gesundheitsförderung – Berichtsjahr 2007. MDS, Essen, S 35–38

Kliche T, Elsholz A, Escher C et al. (2009) Anforderungen an Qualitätssicherungsverfahren für Prävention und Gesundheitsförderung. Eine Expertenbefragung. Präv Gesundheitsf 4 (4): 251–258

Kliche T, Goßmann F (2011a) Was müssen wir über Organisationen wissen, um sie gezielt zu verändern? Flexible qualitative Organisationsdiagnostik mit dem Setting-

Ansatz der WHO. In: Glazinski B, Kramer J (Hrsg) Kairos. Berichte des Instituts für Angewandte Managementforschung, 1/11. Verlag für Angewandte Managementforschung, Köln, S 28–62
Kliche T, Hart D, Kiehl U et al. (2010a) (Wie) wirkt gesundheitsfördernde Schule? Effekte des Kooperationsprojekts »gesund leben lernen«. Präv Gesundheitsf 5 (4): 377–388
Kliche T, Heinrich S, Klein R et al. (2010b) Wirkungsnachweise für die Betriebliche Gesundheitsförderung: Das neue Evaluationssystem der Krankenkassen in Erprobung. Prävention 33 (1): 19–22
Kliche T, Riemann K, Bockermann C et al. (2011b) Gesundheitswirkungen der Prävention: Entwicklung und Erprobung eines Routine-Evaluationssystems für Primärprävention und Gesundheitsförderung der Krankenkassen in Settings, Betrieben und Gesundheitskursen. Gesundheitswesen 73 (4): 247–257
Kliche T, Schreiner-Kürten K, Wanek V et al. (2011c) Gesundheitswirkungen von Prävention: Erprobung des Evaluationssystems der Krankenkassen im Individualansatz und erste Befunde aus 212 Gesundheitskursen. Gesundheitswesen 73 (4): 258–263
Kliche T, Werner AC, Post M (2006) Wie wirkt übergreifende Schulische Gesundheitsförderung? Der Forschungsstand. In: Mittag E, Sticker E, Kuhlmann K (Hrsg) Leistung – Lust und Last. Impulse für eine Schule zwischen Aufbruch und Widerstand. Deutscher Psychologen Verlag, Bonn, S 452–455
Liersch S, Sayed M, Windel I et al. (2012) Innovative Ansätze zum Qualitätsmanagement für eine gesundheitsfördernde Schule. In: BZgA (Hrsg) Gesund aufwachsen in Kita, Schule, Familie und Quartier. Forschung und Praxis der Gesundheitsförderung. BZgA, Köln, S 164–174
Lister-Sharp D, Chapman S, Stewart-Brown S et al. (1999) Health promoting schools and health promotion in schools: two systematic reviews. National Coordinating Centre for Health Technology Assessment NCCHTA, Southampton
Richard JF, Schneider BH, Mallet P (2012) Revisiting the whole-school approach to bullying: Really looking at the whole school. School Psychol Int 33 (3): 263–284
Schempp N, Zelen K, Strippel H (2012) Präventionsbericht 2011. Leistungen der gesetzlichen Krankenversicherung: Primärprävention und betriebliche Gesundheitsförderung. Berichtsjahr 2010. MDS, Spitzenverband Bund der Krankenkassen, Essen, Berlin
Stewart-Brown S (2006) What is the evidence on school health promotion in improving health or preventing disease and, specifically, what is the effectiveness of the health promoting schools approach? WHO Regional Office for Europe Health Evidence Network report. WHO Regional Office for Europe, Copenhagen
Stuppardt R, Wanek V (2009) Qualitätssicherung der primärpräventiven Leistungen der Gesetzlichen Krankenversicherung nach § 20 SGB V. In: Kolip P, Müller V (Hrsg) Qualität von Gesundheitsförderung und Prävention. Huber, Bern, S 177–200
Töppich J, Lehmann H (2009) QIP: Qualität in der Prävention. Ein Verfahren zur kontinuierlichen Qualitätsverbesserung in der Gesundheitsförderung und Prävention. In: Kolip P, Müller V (Hrsg) Qualität von Gesundheitsförderung und Prävention. Huber, Bern, S 223–238
Weare K, Nind M (2011) Mental health promotion and problem prevention in schools: what does the evidence say? Health Promotion Int 26 (suppl 1): i29–i69

Evaluation nationaler Gesundheitsziele in Deutschland

Ulrike Maschewsky-Schneider, Martina Thelen

gesundheitsziele.de ist die Plattform zur Entwicklung nationaler Gesundheitsziele in Deutschland. Unter Beteiligung von Bund, Ländern und Akteuren (der Selbstverwaltung) des Gesundheitswesens entwickelt *gesundheitsziele.de* im Konsens nationale Gesundheitsziele und empfiehlt Maßnahmen zur Zielerreichung. *gesundheitsziele.de* will eine gemeinsame Zielorientierung unterstützen, indem auf Grundlage von Selbstverpflichtungen der Akteure konkrete Maßnahmen zur Zielerreichung umgesetzt werden. Seit Beginn der Arbeiten von *gesundheitsziele.de* im Jahr 2001 ist das Thema Evaluation mitdiskutiert und sukzessive weiterentwickelt worden. Dabei ist nach wie vor sowohl die Entwicklung von Evaluationskonzepten und Evaluation einzelner nationaler Gesundheitsziele als auch die Bewertung zum Nutzen des Gesamtprozesses Gegenstand der Arbeiten.

1 Der Kooperationsverbund *gesundheitsziele.de*

Anliegen von *gesundheitsziele.de* ist die Bildung einer Konsensplattform aller relevanten Akteure im Gesundheitswesen in Deutschland, um nationale Gesundheitsziele zu entwickeln und Gesundheitsziele als Steuerungsinstrument in der Gesundheitspolitik zu etablieren. *gesundheitsziele.de* entwickelt im Konsens Gesundheitsziele, empfiehlt Maßnahmen zur Zielerreichung und stößt Umsetzungsstrategien in Selbstverpflichtung der verantwortlichen Akteure an. Seit dem Jahr 2000 sind mehr als 100 relevante Organisationen des Gesundheitswesens im Kooperationsverbund vereint. Gesundheitsziele in Deutschland haben keine gesetzliche Grundlage und leben daher von den Prinzipien Partizipation und Konsens. Die Bereitschaft der Akteure zum vernetzten Handeln ist die entscheidende Determinante für das Gelingen und den Erfolg der gemeinsamen Zielorientierung (Kooperationsverbund *gesundheitsziele.de* 2010). Die ständigen Gremien von *gesundheitsziele.de* (Ausschuss, Steuerungskreis, Evaluationsbeirat) sowie die Arbeitsgruppen zu den einzelnen Gesundheitszielen bilden eine Plattform für den fachlichen Austausch und die Kooperation mit anderen Akteuren.

In den interdisziplinär besetzten Arbeitsgruppen wurden bisher unter dem Dach von *gesundheitsziele.de* folgende Gesundheitsziele beschlossen:

- Diabetes mellitus Typ 2: Erkrankungsrisiko senken, Erkrankte früh erkennen und behandeln
- Brustkrebs: Mortalität vermindern, Lebensqualität erhöhen
- Tabakkonsum reduzieren
- Gesund aufwachsen: Lebenskompetenz, Bewegung, Ernährung
- Gesundheitliche Kompetenz erhöhen, Patient(inn)ensouveränität stärken
- Depressive Erkrankungen: verhindern, früh erkennen, nachhaltig behandeln
- Gesund älter werden.

Die Auswahl der Ziele erfolgt nach festgelegten Kriterien (s. a. Maschewsky-Schneider et al. 2009). Das sind einerseits wissenschaftliche Kriterien: Diese umfassen die Bestandsaufnahme zum Status der gesundheitlichen Lage der Bevölkerung/Zielgruppe und der bestehenden Versorgung bzw. Prävention (Mortalität, Krankheitslast, Verbreitung, Verbesserungspotenzial bezogen auf Struktur, Versorgung, Finanzierung, Präventionsbedarf, sozialrechtliche und volkswirtschaftliche Relevanz, ethische Aspekte, Chancengleichheit, Beteiligungsmöglichkeiten und Messbarkeit). Andererseits geht es um die Machbarkeit vor dem Hintergrund der gesundheitspolitischen Priorisierungen und der Umsetzbarkeit bzw. Umsetzungsbereitschaft durch die am Zieleprozess beteiligten Akteure. Die Beschreibung und Bewertung des möglichen Zielbereichs mit der Kriterienanalyse bildet eine wichtige Grundlage nicht nur für die Entscheidung für oder gegen dieses Ziel. Mit der Kriterienanalyse erfolgt auch die Bestandsaufnahme der Ausgangslage (Assessment) und sie ist Voraussetzung für die Präzisierung und ggf. Quantifizierung der zu entwickelnden Ziele. Damit bildet sie eine wich-

tige Grundlage für die spätere Evaluation der Zielerreichung.

Gesundheitsziele (Ziele und Teilziele) sollen nach dem SMART-Ansatz entwickelt werden, d. h., dass sie: Specific (sometimes also Simple) (spezifisch/einfach), Measurable (messbar), Achievable (erreichbar), Realistic (realistisch), Timely (terminiert) sein sollen. Die bisher vorliegenden Gesundheitsziele erfüllen je nach Thema und Komplexität diese Anforderungen unterschiedlich und haben häufig einen visionären Charakter (z. B. »Die psychische Gesundheit älterer Menschen ist gestärkt bzw. wiederhergestellt«).

2 Gesundheitsziele erfordern komplexe Interventionen

Die Ziele sind umfassend und komplex. Sie umfassen Outcomes (z. B. Mortalität, Morbidität, Prävalenzen von Risikofaktoren), Impact (z. B. Gesundheitsverhalten, Einstellungen, Lebensweisen, Lebenskompetenz), Determinanten (z. B. soziale Lage, Geschlecht, Wohnen, Arbeit, Familie), Organisationsbedingungen von Maßnahmen (z. B. Gesundheitsförderung im Setting) und Veränderungen/Schaffung von Strukturen für die Zielerreichung (z. B. Gesetzesänderungen, Einführung und Finanzierung neuer Versorgungsstrukturen). Die Ziele umfassen Ober- und Teilziele, Strategien und Maßnahmen zur Zielerreichung und sogenannte Startermaßnahmen. Das sind solche, die besonders bedeutsam sind, erreichbar und zeitnah umsetzbar. Solche zusätzlichen Priorisierungen sind notwendig, um die Akteure zum frühzeitigen Handeln zu motivieren. Strategien und Maßnahmen sollen möglichst evidenzbasiert sein, d. h. auf dem bestmöglichen Wissen zur Erreichbarkeit von Zielen mittels dieser Strategien, Maßnahmen oder Politiken beruhen. Sie können mit Präzisierungen der Zielgruppen, Benennung der wesentlichsten Akteure (Zuständigkeiten, Kooperationspartner) und mit der Forderung nach spezifischen Projekten verbunden sein.

Das Ziel »Gesund aufwachsen: Lebenskompetenz, Bewegung, Ernährung« (Bundesministerium für Gesundheit 2010) ist ein gutes Beispiel für die Komplexität der Ziele und Maßnahmen. Hier werden Ziele in Bezug auf drei Settings (Familie, Kita, Schule) jeweils für die Zielbereiche Lebenskompetenz, Bewegung und Ernährung benannt. Die Teilziele selbst sind auf sehr unterschiedlichen Ebenen angesiedelt. Sie beziehen sich auf Personen der Zielgruppe und ihr Verhalten (z. B. Medienumgang) bzw. Fähigkeiten (z. B. Lebenskompetenz), auf Determinanten (z. B. familiäre Belastungen, Wohnumfeld, Geschlechterbezug), auf Gesundheit (z. B. Behinderung, Entwicklungsverzögerung), auf soziale Netze (z. B. der Familien im Wohnumfeld) und auf Strukturbedingungen, Organisationskulturen (in Schule und Kita) und Aufgaben von Multiplikatoren und Professionellen (z. B. Erzieherinnen, Hebammen).

Für die Umsetzung der Ziele stehen die am Zielprozess beteiligten Akteure, ihre Organisationseinheiten und weitere am Zielprozess nur mittelbar Beteiligte. Sie richten ihr professionelles und institutionelles Handeln an den gemeinsam vereinbarten Zielen aus, weil sie von deren Relevanz für die Gesundheit ihrer Zielgruppen überzeugt sind. Die Umsetzung erfolgt unter den Alltagsbedingungen ihres Handelns, d. h., es wird an die in Deutschland bislang bestehenden Strukturen und Aktivitäten in den Zielbereichen angeknüpft und diese werden durch weitergehende oder verstärkende Maßnahmen ergänzt. Die Auswahl und Umsetzung von Maßnahmen und Programmen ist damit an den jeweiligen institutionellen Kontext des Akteurs gebunden (z. B. an die gesetzgeberische Kompetenz des Bundes, an die durch den Träger vorgegebene Organisationskultur einer Kita, an die professionelle Orientierung des Kinderarztes in der ärztlichen Versorgung, an die Finanzierungsmodalitäten der ambulanten Versorgung durch Hebammen).

Die Ziele erfüllen damit alle relevanten Kriterien von komplexen Interventionen (Craig et al. 2008; Milton et al. 2010). Es handelt sich bei den Gesundheitszielen um Interventionen unter Alltagsbedingungen, wobei die Interventionen nicht einfach umrissene Maßnahmen sind (z. B. die Teilnahme an einem Ernährungskurs), sondern es geht um komplexe Politiken und Strategien bezogen auf Bevölkerungen (oder Subgruppen) und bezogen auf unterschiedliche strukturelle Ebenen (Bund, Länder, Kommunen, Organisationen, Institutionen). Das bedeutet, dass die verschiedenen Komponenten der Intervention miteinander in vielfältiger Weise interagieren (Craig et al. 2008, S. 7). Die Ziele umfassen mehrere Outcomes auf sehr

unterschiedlichen Ebenen und beziehen sich auf verschiedene Zielgruppen. Die Interventionen sind nicht standardisiert und können sich sogar im Interventionsverlauf verändern (ebd). Stärke und Umfang der Intervention können je nach Handlungs- und Zielbereich sehr unterschiedlich sein und sind im Hinblick auf die mögliche Messbarkeit nicht vergleichbar. Zwar ist die Relevanz der Interventionen aufgrund der Alltagsbedingungen, in denen sie verläuft, und die Transferierbarkeit auf andere Alltagssituationen hoch (z. B. bewährte Gesundheitsziele in einem Bundesland können gut auf ein anderes übertragen werden, da die Kontexte vergleichbar sind) (externe Validität), die Komplexität der Intervention erschwert jedoch die Möglichkeit, Wirkungen bei den angestrebten Gesundheitsindikatoren auf die Intervention als Ganzes oder von Teilinterventionen zurückführen zu können (interne Validität).

Für die Evaluation von Gesundheitszielen als komplexe Programme und Politiken ergeben sich damit grundlegende methodische Probleme und Begründungszusammenhänge. Milton et al. (2010) diskutieren neun solcher Herausforderungen, die sich auch auf die Evaluation von Gesundheitszielen gut anwenden lassen. Sie beziehen sich auf das Problem fehlender Vergleichsgruppen, lange Latenzzeiten, bis Interventionen auf Gesundheit wirken, das Problem der Messung von sozialen Einflussfaktoren außerhalb des Gesundheitswesens, die Schwierigkeit, bei multimodalen und auf verschiedenen Ebenen angelegten Interventionen Ursache-Wirkungszusammenhänge zu analysieren, die Kontextgebundenheit von Interventionen und die Spannung zwischen der Forderung der Politik nach schnellen Ergebnissen und den eher langwierigen wissenschaftlich-methodischen Forschungsprozessen (s. a. Tab. 1, Spalte 1).

Tabelle 1
Methodische Herausforderungen an die Evaluation von Gesundheitszielen
Quelle: nach Milton et al. 2010

Methodische Herausforderungen	*gesundheitsziele.de*
Problem der Vergleichsgruppen (Kontrollen)	
Das Ziel wird unter Alltagsbedingungen im ›normalen‹ gesellschaftlichen, sozialen, politischen und Versorgungskontext umgesetzt. Vergleichsgruppen sind schwer zu konstruieren.	Die Umsetzung von Gesundheitszielen in Deutschland ist an den Kontext des deutschen Gesundheits- und Sozialsystems gebunden. Ein Vergleich mit anderen Ländern ist in Bezug auf einzelne Indikatoren sinnvoll (z. B. Auswirkungen der Tabaksteuererhöhung auf das Rauchverhalten von jungen Menschen), aber nicht vorrangiges Ziel der Evaluation.
'Time-lags' und lange Kausalketten	
Ergebnisse in Bezug auf Gesundheit werden erst nach längerer Interventionszeit erzielt.Es bestehen komplexe Kausalketten, die über verschiedene soziale Mechanismen und Determinanten auf die Gesundheit wirken. Hierzu können vor dem Hintergrund des empirischen und theoretischen Wissens Annahmen zu Wirkungsketten entwickelt und intermediäre Messindikatoren bestimmt werden.	Für viele der Teilziele bei *gesundheitsziele.de* werden Ergebnisse erst mit Zeitverzögerung erreicht (z. B. Auswirkungen der Interventionen zur Lebenskompetenz bei Kindern auf das lebenslange Gesundheitsverhalten und die Gesundheit selbst). Mit der Formulierung von Teilzielen und Startermaßnahmen können jedoch Modelle gebildet werden, die zeigen, über welche Interventionen und Zwischenschritte Lebenskompetenz verbessert werden kann. Für diese lassen sich dann passende Messindikatoren bilden (z. B. Bestandsaufnahme zu schulischen Programmen zur Gewaltprävention in Deutschland).
Politische Maßnahmen mit Gesundheitsindikatoren verknüpfen	
Politische Maßnahmen (Interventionen), die sich auf soziale Determinanten beziehen, die auch außerhalb des Gesundheitswesens liegen können, lassen sich in ihren Wirkungen auf Gesundheitsindikatoren nur schwer messen. Effekte werden möglicherweise unterschätzt. Die Autoren empfehlen die Ableitung von intermediären Indikatoren auf unterschiedlichen Messebenen. Auch an die Evaluation im Rahmen von Fallstudien wäre zu denken.	Dies trifft auch auf die Ziele von *gesundheitsziele.de* zu. Im Ziel »Gesund aufwachsen« werden im Ziel 10 optimierte Rahmenbedingungen und Strukturen für Gesundheitsförderung in den drei Settings gefordert. Ein Unterziel fordert die bessere Vernetzung der nach SGB zuständigen Akteure aus dem Gesundheits- und aus dem Sozialbereich. Dieses Ziel stellt an die Evaluation enorme Anforderungen. Können Vernetzungsformen auf kommunaler Ebene über den Einzelfall hinaus überhaupt flächendeckend für Deutschland evaluiert werden? Fallstudien in einzelnen Kommunen könnten angemessen sein, um aus deren Ergebnissen Prognosen zu Langzeiteffekten abzuleiten.

Methodische Herausforderungen	*gesundheitsziele.de*
Umgang mit Komplexität	
Bei komplexen Interventionen sind die verschiedenen Einflussfaktoren und -ebenen schwer zu prüfen. Notwendig ist deshalb die theoretische Fundierung von Intervention und Evaluation und die Entwicklung von Kausalmodellen und Wirkungsketten. Auf dieser Grundlage können relevante und messbare Indikatoren auf den verschiedenen Ebenen identifiziert werden.	In die Zielentwicklung gehen theoretische Grundlagen ein. Im Gesundheitsziel »Gesund aufwachsen« sind diese explizit ausgeführt und wissenschaftlich begründet. Für die Zielentwicklung sind die beiden theoretischen Konzepte Lebenskompetenz und Intervention im Setting handlungsleitend. Auf dieser Grundlage können für ein Evaluationskonzept mögliche Wirkungsketten antizipiert und relevante Indikatoren definiert werden.
Übertragbarkeit und die Bedeutung des Kontextes	
Ergebnisse auf Outcome-Ebene (Gesundheitsindikatoren) werden vor dem Hintergrund des jeweiligen politischen und Versorgungskontexts erzielt. Jede Region, Organisation und jedes Setting hat einen spezifischen sozialen, regulativen und historischen Kontext, der die jeweiligen Ergebnisse determiniert. Welche Einflüsse damit verbunden sind und wie Kontextdeterminanten und spezifische Interventionen interagieren, ist oft nicht bekannt und nur schwer messbar. Damit stellt sich die Frage der Übertragbarkeit der Interventionen und Ergebnisse auf andere Kontexte.	*gesundheitsziele.de* hat nicht den Anspruch, Ziele und Interventionskonzepte auf andere Kontexte außerhalb Deutschlands (z. B. andere Länder) zu übertragen. Vielmehr geht es um den spezifischen deutschen Kontext. Andererseits ist es wichtig zu wissen, welche Interventionen und Politiken wirksam sind, um von den Erfahrungen anderer zu profitieren oder Politiken aufeinander abzustimmen (z. B. Strukturmaßnahmen/gesetzliche Regelungen zur Tabakkontrollpolitik auf EU-Ebene). Die Übertragbarkeit der Evidenzlage auf den jeweiligen Kontext (z. B. von einem Bundesland auf ein anderes) ist also zu prüfen. Entscheidungen hinsichtlich der Übertragung von in anderen Kontexten erfolgreich durchgeführten Interventionen erfordern neben wissenschaftlichen auch immer soziale, politische und ethische Begründungen.
Ungleichgewicht zwischen wissenschaftlichen und politischen Zeitfenstern	
Politik und Wissenschaft folgenden verschiedenen Zeitregimes. Politik braucht schnelles Handeln und zeitnahe Entscheidungsgrundlagen. Wissenschaft folgt wissenschaftlichen Standards und Kriterien, die gründliche Prüfung und methodische Sorgfalt erfordern. Evaluation von Public Health-Maßnahmen erfolgt im Spannungsfeld dieser verschiedenen Anforderungen.	*gesundheitsziele.de* hat für dieses Dilemma eine pragmatische Lösung gefunden. Ein Teil der geforderten Aufgaben zur Zielentwicklung (Bestandsaufnahme nach der Kriterienanalyse, Prüfung der wissenschaftlichen Evidenz) erfolgt im Vorfeld der Zielentwicklung bzw. in den AGs zur Zielentwicklung. Die Orientierung an den SMART-Kriterien ist eine gute Vorbereitung für die Entwicklung von Evaluationskonzepten. Durch die Priorisierung auf die wichtigsten Ziele (Startermaßnahmen), wird die Entwicklung eines Evaluationskonzepts vereinfacht. Die Entscheidung, die Evaluation möglichst auf der Grundlage bestehender Datenquellen und mit Fokus auf die nationale und Bevölkerungsebene durchzuführen, fördert ebenfalls die zügige Erstellung von Evaluationskonzepten und deren Durchführung. Auch sind die Evaluationskonzepte flexibel und offen und lassen sich an veränderte Rahmenbedingungen für die Zielerreichung gut anpassen.
Die Schwierigkeit, Entscheidungsträger aus anderen Sektoren einzubinden	
Da viele Interventionen, die für die Gesundheit der Menschen relevant sind, außerhalb des Gesundheitswesens stattfinden, sollten die entsprechenden Akteure auch einbezogen werden. Diese mögen über die relevanten Daten verfügen.	Bei der Entwicklung der Ziele werden auch die relevanten Akteure aus anderen Sektoren einbezogen (z. B. beim Ziel »Gesund aufwachsen« die verschiedenen für die Zielgruppen zuständigen Ministerien auf Bundes- und Landesebene). Das erleichtert die Evaluation, da diese Ministerien über notwendige Daten verfügen (z. B. Statistisches Bundesamt, Rentenversicherungsträger).
Priorisierung von Interventionen für die Evaluation	
Vor dem Hintergrund, dass nur Ausschnitte von komplexen Interventionen evaluierbar sind, sollte Wissenschaft sich auf die Evaluation von solchen Interventionen konzentrieren, aus denen wichtige Erkenntnisse für politische Entscheidungen erzielt werden können.	Bei *gesundheitsziele.de* wurden in den AGs bereits Priorisierungen für die Ziele vorgenommen. Die Evaluation bzw. Evaluationskonzepte fokussieren ausschließlich auf diese ›Startermaßnahmen‹ (z. B. beim Ziel »Tabakkonsum kontrollieren« auf strukturelle Maßnahmen und gut messbare Prävalenzen).

Methodische Herausforderungen *gesundheitsziele.de*

Spannung zwischen ›robuster‹ und ›ausreichender‹ Evidenz

Evaluation soll Wissen schaffen, das Entscheidungsträger davon überzeugt, dass Interventionen wirksam und nützlich sind, um Nachhaltigkeit der Politiken sicherzustellen. Dies erfordert eine Fokussierung der Evaluation, die den Erfordernissen politischer und gesellschaftlicher Entscheidungslogiken folgt und nicht wissenschaftlichen Partialinteressen. Das bedeutet, Evaluationen auf wesentliche entscheidungsrelevante Indikatoren zu konzentrieren.

Die Evaluationskonzepte von *gesundheitsziele.de* sind wissenschaftlich sehr gut fundiert, sie folgen aber dem pragmatischen Ziel der Konzentration auf das Wesentliche. Es sollen solche Informationen auf allen drei Messebenen: Struktur, Ergebnis und Prozess gesammelt und verwendet werden, die für politische Entscheidungen relevant sind, ggf. Entscheidungen provozieren oder einen politischen Diskurs und Kontroversen anregen.

3 Evaluation nationaler Gesundheitsziele

Evaluation ist untrennbarer Bestandteil von Gesundheitszielprogrammen, so auch von *gesundheitsziele.de*. Aufgabe der Evaluation ist es, die Zielerreichung wissenschaftlich zu belegen. Da in die Entwicklung und vor allem in die Umsetzung von Gesundheitszielen nicht unbeträchtliche öffentliche Ressourcen fließen, ist Transparenz über die Ergebnisse dieses Prozesses geboten. Die Evaluation kann auf verschiedenen Ebenen erfolgen. Das ist der Prozess der Konsensfindung zwischen den Akteuren, die Zielerreichung auf Ergebnis-, Prozess- und Strukturebene und die Handlungs- und Wertorientierung der Beteiligten. Tabelle 2 gibt einen Überblick über die verschiedenen Ebenen.

Tabelle 2
Funktionen des Zielprozesses und Evaluationsebenen
Quelle: nach Maschewsky-Schneider 2005

Funktionen des Zieleprozesses	Evaluationsebenen
Konsensbildung	Evaluation des Zielfindungsprozesses
Ergebnisorientierung	Evaluation der messbaren Zielerreichung mittels Gesundheitsindikatoren oder abgeleiteter intermediärer Indikatoren
Strukturbildung	Evaluation der strukturellen Interventionen im Gesundheitswesen
Maßnahmenorientierung	Evaluation von Einzelmaßnahmen und Kampagnen
Handlungsorientierung	Evaluation des Gesamtprozesses, einschließlich der zugrunde liegenden Wertorientierungen und der sozialen und ethischen Prinzipien des Handelns

Um die Fragen und Aufgaben hinsichtlich der Evaluation kontinuierlich zu begleiten, wurde im Jahr 2004 der Evaluationsbeirat eingerichtet. Seine Aufgaben sind die wissenschaftliche Begleitung des Kooperationsverbundes, die Evaluation des Gesamtprozesses und der Gesundheitsziele. Ziel der Evaluation von nationalen Gesundheitszielen ist die Abbildung des Zielerreichungsgrades und der Umsetzung von Maßnahmen. Die Ergebnisse der Evaluation ermöglichen eine effektivere und effizientere Nachsteuerung des jeweiligen Gesundheitsziels (GVG 2008).

Der Evaluationsbeirat einigte sich darauf, den Fokus der Zielevaluation auf die Ergebnisevaluation zu legen und hierzu möglichst vorliegende, repräsentative Datenquellen, insbesondere aus der Gesundheitsberichterstattung zu nutzen (s. a. Maschewsky-Schneider 2005). An die Arbeitsgruppen, die die Ziele entwickelten, wurde die Anforderung gestellt, diese so eindeutig zu formulieren, dass daraus messbare Evaluationsindikatoren ableitbar sind. Vor dem Hintergrund verschiedenster Gesetzgebungsverfahren (z. B. in Bezug auf neue Versorgungsmodelle oder Besteuerung von Zigaretten), der Entwicklung von Gesundheitszielen in den Bundesländern und dem Ausbau der Prävention und Gesundheitsförderung bei den Krankenkassen wurde beschlossen, für die Evaluation auch die wichtigsten strukturellen Änderungen und Rahmenbedingungen mit zu dokumentieren. Schwierig stellte sich aus der Sicht des Evaluationsbeirats jedoch die Evaluation auf der Maßnahmenebene dar. Viele Akteure in Deutschland tragen mit ihren Aktivitäten zur Zielerreichung bei, eine Übersicht zu erstellen oder den Beitrag der Einzelmaßnahmen umfassend zu evaluieren, schien aussichtslos.

Es wurde deshalb beschlossen, genau zu prüfen, welche Prozessindikatoren gefunden werden könnten. Es sollten möglichst nur deutschlandweit ausgerichtete Maßnahmen einbezogen werden. Für sie sollten Datenquellen bereits zur Verfügung stehen oder sie sollten so entscheidend sein, dass für sie Daten verfügbar gemacht werden müssten. Zu solchen relevanten Maßnahmen gehören insbesondere groß angelegte Kampagnen, z. B. die der Bundeszentrale für gesundheitliche Aufklärung, für die bereits ein Monitoringsystem vorliegt und damit entscheidende Informationen für die Evaluation bereitstellt.

Die Evaluation soll Schlussfolgerungen erlauben, warum die Ziele erreicht oder nicht erreicht wurden, um ggf. programmbegleitend Verbesserungen einzuleiten. Dieser Ansatz setzt ein Wirkmodell voraus, d. h. Hypothesen über den Beitrag von Maßnahmen zur Erreichung von Zielen und die Benennung der Determinanten für Erfolg oder Misserfolg. Die Ergebnisse der Evaluation nationaler Gesundheitsziele werden an die beteiligten Akteure und umsetzenden Einrichtungen zurückgekoppelt mit dem Ziel, Qualitätsverbesserungen bei der Maßnahmendurchführung anzustoßen und Ziele zu aktualisieren.

4 Evaluationskonzepte für die Zielevaluation

Grundlage für die Evaluation nationaler Gesundheitsziele sind Evaluationskonzepte, die vom Evaluationsbeirat unter Mitarbeit von wissenschaftlichen Experten und Expertinnen und der entscheidenden Datenhalter erarbeitet werden. Für drei der sieben Gesundheitsziele liegen Evaluationskonzepte vor:

- »Tabakkonsum reduzieren« (Kröger et al. 2010)
- »Depressive Erkrankungen: verhindern, früh erkennen, nachhaltig behandeln« (Bermejo et al. 2009)
- »Patient(inn)ensouveränität stärken« (Horch et al. 2009).

Die Evaluationskonzepte stehen unter *www.gesundheitsziele.de* als Download zur Verfügung. Bei der Entwicklung der Evaluationskonzepte sprach der Beirat sich für ein pragmatisches Konzept aus, das einfach handhabbar, finanzierbar und für die Beteiligten und Betroffenen akzeptabel und machbar ist. Im Rahmen der Evaluationskonzepte wurden nach einer festgelegten Systematik Indikatorensysteme aufgestellt, die im Sinne eines Monitorings periodische Veränderungen auf Bevölkerungsebene abbilden können. Neben der Entwicklung geeigneter Indikatoren wurden ferner Datenquellen und -lücken und die jeweiligen Datenhalter identifiziert.

gesundheitsziele.de hat im Jahr 2003 das nationale Gesundheitsziel »Tabakkonsum reduzieren« erstmals vorgelegt, das die vier Handlungsfelder »Effektive Tabakkontrollpolitik«, »Förderung des Ausstiegs«, »Verhinderung des Einstiegs« und »Schutz vor Passivrauchen« (BMG 2003) umfasst. Auf Grundlage des Evaluationskonzeptes (Maschewsky-Schneider et al. 2006) und bestehender Datenbanken wurde 2009 eine Evaluation der Startermaßnahmen: (1) Tabaksteuererhöhung, (2) vollständiges Verbot direkter und indirekter Tabakwerbung, (3) Schutz vor Passivrauchen, (4) Förderung des Ausstiegs und (5) Maßnahmen zur Verhinderung des Einstiegs in das Rauchen durchgeführt. Die Evaluation zeigt, dass eine Reduzierung der Raucherprävalenzen bei Jugendlichen erreicht werden konnte (Kröger et al. 2010). Durch die Priorisierung mittels der vier Startermaßnahmen und die Konzentration der Evaluation auf diese lassen sich begründete Schlussfolgerungen hinsichtlich der Wirkungszusammenhänge ziehen. Evidenzbasierte strukturelle Maßnahmen (Steuererhöhung, Rauchverbot in öffentlichen Räumen, Schulen, Restaurants und Clubs) und gestiegene Aufmerksamkeit für das Nichtrauchen durch deutschlandweit angelegte Kampagnen und Programme konnten vermutlich diesen Effekt erzielen (s. a. Bundesgesundheitsblatt 2010, Themenheft »Tabakprävention in Deutschland: Maßnahmen und Erfolge«).

Die Ergebnisse belegen auch, dass weiterhin Handlungsbedarf in der Tabakprävention besonders in Hinblick auf die Nachhaltigkeit der Interventionen gegeben ist. Die Evaluationsergebnisse sind Grundlage für die Arbeitsgruppe, das Gesundheitsziel »Tabakkonsum reduzieren« zu aktualisieren und damit neue Akzente für die Umsetzung von Maßnahmen zu setzen.

5 Evaluation des Gesamtprozesses

Von besonderem Interesse ist, ob die Gesundheitsziele sich in konkretes Handeln der Akteure umsetzen bzw. ob Maßnahmen ergriffen werden, die zur Zielerreichung beitragen. In den Jahren 2007 und 2008 hat die Geschäftsstelle daher umfangreiche Befragungen der am Gesundheitszieleprozess beteiligten Akteure durchgeführt. Erfasst wurden die konkreten Maßnahmen der beteiligten Organisationen, die einen Beitrag zur Erreichung der Gesundheitsziele leisten. Aufbauend auf den Befragungen empfiehlt der Evaluationsbeirat eine retrospektive qualitative Erhebung bei den Akteuren. Dabei soll der Durchdringungsgrad des Zielegedankens bei den beteiligten Kooperationspartnern erfasst und der Nutzen des nationalen Gesundheitszieleprozesses bewertet werden. Damit wird die Ebene der Evaluation der Ziele und Teilziele ergänzt um die der Evaluation der Handlungs- und Politikorientierung (Tab. 1).

6 Schlussfolgerungen

Es konnte gezeigt werden, dass es sich bei den Gesundheitszielen um komplexe Interventionen handelt, für deren Evaluation ein einfaches, flexibles und kostengünstiges Konzept entwickelt wurde, das sich an dem Kriterium der Relevanz für politische Entscheidungsfindung orientiert. Geht man zurück auf die genannten methodischen Herausforderungen für die Evaluation von komplexen Interventionen und bezieht diese auf *gesundheitsziele.de*, dann zeigt sich, dass dafür gute und praktikable Lösungen gefunden werden konnten. (Tab. 1, Spalte 2).

Ein Anspruch auf Verallgemeinerung und den Vergleich der Evaluationsergebnisse mit anderen Ländern besteht nur punktuell hinsichtlich spezifischer Indikatoren und Maßnahmen. Durch die Formulierung von Teilzielen und Startermaßnahmen und auf der Grundlage von Theorien und wissenschaftlicher Evidenz können Modelle zu Wirkungsketten entwickelt und auch bei langen Latenzzeiten mittels intermediärer Wirkindikatoren gemessen werden. Da auch Akteure aus anderen Bereichen als dem Gesundheitswesen eingebunden sind, kann nicht nur auf deren ›Interventionen‹ gebaut werden, sondern auch deren Daten und Erkenntnisse können für die Evaluation mit verwendet werden. Das Evaluations- und Interventionskonzept sind pragmatisch und ergebnisorientiert. Die Formulierung von Startermaßnahmen und die Konzentration der Evaluationen auf diese ist erfolgversprechend und verhindert, dass Evaluation in Verästelungen wissenschaftlichen Handelns stecken bleibt, sondern zügig handlungsrelevantes Wissen schafft.

Die Evaluationskonzepte von *gesundheitsziele.de* sind wissenschaftlich fundiert, sie folgen dem pragmatischen Ziel der Konzentration auf das Wesentliche. Es werden Erkenntnisse auf Struktur-, Ergebnis- und Prozessebene geschaffen, die für politische Entscheidungen relevant sind, Entscheidungen provozieren und einen politischen Diskurs und Kontroversen anregen. Sowohl den Zielen als auch den Evaluationskonzepten liegen neben den wissenschaftlichen ethische und soziale Wertorientierungen zugrunde, die sich am Kriterium des Ausgleichs sozial bedingter gesundheitlicher Ungleichheit ausrichten.

Literatur

Bermejo I, Klärs G, Böhm K et al. (2009) Evaluation des nationalen Gesundheitsziels »Depressive Erkrankungen: verhindern, früh erkennen, nachhaltig behandeln«. Bundesgesundheitsbl – Gesundheitsforsch – Gesundheitsschutz 52: 897–904

Bundesministerium für Gesundheit (BMG) (Hrsg) (2003) gesundheitsziele.de – Forum zur Entwicklung und Umsetzung von Gesundheitszielen in Deutschland. Bericht. Tabakkonsum reduzieren. Berlin

Bundesministerium für Gesundheit (Hrsg) (2010) Nationales Gesundheitsziel: Gesund aufwachsen: Lebenskompetenz, Bewegung, Ernährung. *gesundheitsziele.de* Kooperationsverbund zur Weiterentwicklung des nationalen Gesundheitszieleprozesses

Bundesgesundheitsblatt – Gesundheitsforschung – Gesundheitsschutz (2010) 53. Springer-Verlag. Themenheft »Tabakprävention in Deutschland: Maßnahmen und Erfolge«

Craig P, Dieppe P, Macintyre S et al. (2008) Prepared on behalf of the Medical research Council. Developing and evaluating complex interventions: new guidance www.mrc.ac.uk/complexinterventionsguidance

Gesellschaft für Versicherungswissenschaft und -gestaltung e. V. (GVG) (2008) Ziele auswählen, entwickeln und evaluieren. Zentrale Konzepte von *gesundheitsziele.de*

Horch K, Hölling G, Klärs G et al. (2009) Ansätze zur Evaluation des Gesundheitsziels »Gesundheitliche Kompetenz erhöhen, Patient(inn)ensouveränität stärken«. Bundesgesundheitsbl – Gesundheitsforsch – Gesundheitsschutz 52: 889–896

Kooperationsverbund *gesundheitsziele.de* (2010) Gemeinsame Erklärung des Kooperationsverbundes zur Weiterentwicklung des nationalen Gesundheitszieleprozesses http://www.*gesundheitsziele.de* (Stand: 20.04.2012)

Kröger C, Mons U, Klärs G et al. (2010) Evaluation des Gesundheitsziels Tabakkonsum. Bundesgegesundheitsbl – Gesundheitsforsch – Gesundheitsschutz 53: 91–102

Maschewsky-Schneider U (2005) Programm-, Ziel- und Maßnahmenevaluation bei *gesundheitsziele.de*. In: Gesellschaft für Versichungswirtschaft e. V. (Hrsg) *gesundheitsziele.de* – Impulse, Wirkungen und Erfahrungen. Akademische Verlagsgesellschaft, Berlin

Maschewsky-Schneider U, Lampert T, Kröger C et al. (2006) Evaluation des Gesundheitsziels »Tabakkonsum reduzieren«. Bundesgesundheitsbl – Gesundheitsforsch – Gesundheitsschutz 49: 1155–1161

Maschewsky-Schneider U, Klärs G, Ryl L et al. (2009) *gesundheitsziele.de*. Ergebnisse der Kriterienanalyse für die Auswahl eines neuen Gesundheitsziels in Deutschland. Bundesgesundheitsbl – Gesundheitsforsch – Gesundheitsschutz 52: 764–774

Milton B, Moonan M, Taylor-Robinson D et al. (eds) (2010) How can health equity impact of universal policies be evaluated? Insights into new approaches and next steps. Liverpool WHO Collaborating Centre for Policy Research on Social Determinants of Health

Evaluation der Gemeinsamen Deutschen Arbeitsschutzstrategie

Britta Schmitt

1 Ausgangssituation

Die Gemeinsame Deutsche Arbeitsschutzstrategie (GDA) ist die von Bund, Ländern und Unfallversicherungsträgern (UVT) gemeinsam getragene, bundesweit geltende Arbeitsschutzstrategie der Bundesrepublik Deutschland. Inspiriert durch den Occupational Safety and Health (OSH-)Strategy Process der EU (seit 2007) hat die deutsche Arbeitsschutzstrategie (seit 2008) das Ziel, »Sicherheit und Gesundheit der Beschäftigten (...) durch einen präventiven und systemorientierten betrieblichen Arbeitsschutz ergänzt durch Maßnahmen einer betrieblichen Gesundheitsförderung zu erhalten, zu verbessern und zu fördern« (GDA 2007, S. 3). Auf Basis einer engen Abstimmung und durch gemeinsam festgelegte nationale Arbeitsschutzziele sollen auch die Zusammenarbeit der Aufsichtsdienste der Arbeitsschutzbehörden der Länder und der Unfallversicherungsträger bei deren Beratungs- und Überwachungstätigkeit in den Betrieben verbessert sowie die Neuordnung des Vorschriften- und Regelwerks im Sinne größerer Kohärenz, Konsistenz und Nutzerfreundlichkeit vorangetrieben werden.

Da die Gemeinsame Deutsche Arbeitsschutzstrategie (GDA) mehrere – in der ersten Strategieperiode 2008–2012 insgesamt elf – sogenannte Arbeitsprogramme unter dem Dach einer gemeinsamen strategischen Zielsetzung bündelt, zu deren Umsetzung eine Vielzahl von Akteurinnen und Akteuren, insbesondere aus den Reihen der GDA-Träger, d. h. von Bund, Ländern und UVT, aber auch aus den Reihen der Sozialpartner, der Wissenschaft, der Krankenkassen und anderer Verbände und Organisationen für Sicherheit und Gesundheit bei der Arbeit beitragen, kann die GDA geradezu als ein Paradebeispiel für eine komplexe Intervention im Sinne des in diesem Band verwendeten Komplexitätsbegriffs gelten. Dies gilt umso mehr, als die Umsetzung der GDA auf unterschiedlichen Interventionsebenen und zu verschiedenen Themen, in verschiedenen Branchen und Tätigkeitsfeldern geschieht. In der gegenwärtigen Periode tragen die vielfältigen Programme zu den drei nationalen Arbeitsschutzzielen Reduktion von Arbeitsunfällen (kurz: »Arbeitsunfälle«), Reduktion von arbeitsbedingten Muskel-Skelett-Erkrankungen (kurz: »MSE«) und Reduktion von arbeitsbedingten Hauterkrankungen (kurz: »Haut«) bei.

Die erwähnten elf GDA-Arbeitsprogramme verfolgen dabei sehr unterschiedliche Ansätze – vom Online-Selbstbewertungstool für Betriebe der Pflegebranche über Schwerpunktsetzungen in der klassischen Überwachungs- und Beratungsarbeit u. a. bei Zeitarbeit, Bau, Transport und Feuchtarbeit (Gefährdung der Haut) bis hin zu einer eher exemplarischen Präventionsberatung für eine kleinere Zahl von Betrieben beispielsweise aus dem Bereich Feinmechanik, Ernährungsindustrie oder Hotellerie- und Gastronomie etc.

Die drittelparitätisch (UVT, Länder, BMAS) besetzten Arbeitsprogrammgruppen übernehmen die separaten Evaluationen dieser als Einzelmaßnahmen zu verstehenden Arbeitsprogramme selbst. Zum Teil lassen sie sich dabei durch externe Evaluationsprofis oder Datentreuhänder unterstützen. Gleichzeitig liefern sie Beiträge zu den – in allen Arbeitsprogrammen einheitlich durchzuführenden – »Kopfbogen-Erhebungen« und zu den konsensual verabredeten, arbeitsprogrammübergreifenden Indikatoren für die Erreichung der drei o. g. nationalen Arbeitsschutzziele der ersten GDA-Periode. Damit generieren sie wesentliche Anhaltspunkte für die Ergebniszusammenfassung im Rahmen der Gesamtevaluation der Strategie, der sogenannten GDA-Dachevaluation. Diese zielt im Bereich der Arbeitsprogramme vorrangig auf eine Bewertung der Frage ab, inwieweit die nationalen Arbeitsschutzziele der ersten GDA-Periode branchen- und tätigkeitsübergreifend erreicht wurden. Mit dem Abschlussbericht der GDA-Dachevaluation, die im Auftrag der Bundesanstalt für Arbeitsschutz und Arbeitsmedizin (BAuA) von der Kooperationsstelle Hamburg IFE GmbH im Verbund mit TNS Infratest durchgeführt wird, ist Ende 2013 zu rechnen.

2 Das Zielebenen-Konzept der GDA-Dachevaluation

Die Endpunkte der unter dem Dach der GDA gebündelten Interventionen werden nicht nur in den avisierten Verbesserungen der gesundheitlichen Situation der Erwerbsbevölkerung sowie in einer Steigerung der Zahl von Betrieben mit systematischer Arbeitsschutzorganisation (Zielebene 1) gesehen. Mindestens ebenso wichtige Endpunkte liegen in einer verbesserten Zusammenarbeit und verbesserter gegenseitiger Information der Strategie-Akteure selbst (Zielebene 3), womit in erster Linie die Zusammenarbeit der Länder und UVT bei deren Überwachungs- und Beratungstätigkeit in den Betrieben gemeint ist, aber auch die ebenfalls zu verbessernde Kooperation der GDA-Träger mit »externen« Kooperationspartnern (Zielebene 4), allen voran mit den Krankenkassen.

Diese kooperationsbezogenen Zielebenen tragen wesentlich zur Komplexitätssteigerung der Strategie und damit auch zur notwendigen Differenziertheit ihrer Evaluierung bei. Gleiches gilt für die angestrebten Verbesserungen im legislativen Bereich. Da bislang sowohl der Bund im Hinblick auf die maßgeblichen Bundesgesetze, als auch die Bundesländer mit ihren Durchführungs-Verordnungen sowie die UVT bezüglich besonderer Branchenregelungen über relevante Rechtsetzungskompetenzen verfügen, erstreckt sich die GDA im Interesse der Betriebe auch auf das Ziel, ein kohärentes, aufeinander abgestimmtes und möglichst widerspruchsfreies Vorschriften- und Regelwerk im Arbeitsschutz zu schaffen (Zielebene 2: »Akzeptanz und Wirksamkeit des institutionellen Arbeitsschutzes«). Auch in diesem Bereich sind also Evaluationen zu ermöglichen, was in erster Linie durch das sogenannte »Betriebsbarometer«, einer perspektivisch in jeder Strategieperiode zu wiederholenden repräsentativen Befragung von 6.500 Inhabern, Geschäftsführern oder leitenden Angestellten mit Zuständigkeit für den Arbeitsschutz geschehen soll. Die Respondentinnen und Respondenten werden dabei – neben vielem anderen – zu ihrem Kenntnisstand sowie zu ihrer Einschätzung bzgl. der Überschaubarkeit und Praxistauglichkeit des Vorschriften- und Regelwerks befragt. Die erste Befragungswelle hierzu erfolgte bereits von Mai bis August 2011. Begleitend wurden von TNS Infratest auch 5.500 repräsentativ für die Branchenstruktur der Bundesrepublik ausgewählte Beschäftigte zum Arbeitsschutz-Engagement und zur Präventionskultur ihres Arbeitgebers sowie zu ihrem individuellen Arbeitsschutzengagement und ihrer persönlichen Gesundheitskompetenz befragt. Vorläufige Ergebnisse aus den beiden aufeinander bezogenen Befragungen wurden im zweiten Quartal 2012 veröffentlicht.

Sind bis hierhin alle genannten Zielebenen 1–4 noch mit attribuierten Arbeitsprogrammen oder anderen Aktivitäten – etwa der Entwicklung von gemeinsamen Handlungs-Leitlinien für die Aufsichtsdienste, mit organisierten Erfahrungsaustauschen und regelmäßigen Kooperationspartnergesprächen oder der Erarbeitung des im August 2011 verabschiedeten Leitlinienpapiers zum Vorschriften- und Regelwerk etc. – hinterlegt, so widmet sich Zielebene 5 ausschließlich der Beschreibung sehr langfristiger und nur sehr indirekt mit der GDA in Verbindung zu bringender Impacts wie beispielsweise der Steigerung der betrieblichen Wettbewerbsfähigkeit oder der indirekt möglicherweise zu erzielenden (Teil-)Entlastung der sozialen Sicherungssysteme.

Wie diese kurze Skizze der relevanten Zielebenen illustriert, wird die »Dachevaluation« der ersten GDA-Periode auf Grundlage einer Konzeption durchgeführt, deren wesentliche Elemente die erwähnten fünf Zielebenen, fachliche, administrative und organisatorische Rahmenbedingungen sowie "Terms of Reference (ToR)" zur Definition der Ziele, Maßnahmen, Indikatoren, Erhebungsinstrumente und Meilensteine waren und sind. Das Evaluationskonzept der »komplexen Intervention GDA« folgte und folgt in der ersten Periode also dem klassischen Ansatz, der sich im Wesentlichen auf einen zielorientierten Soll-Ist-Vergleich bzw. auf eine erste Nullmessung relevanter Zielbereiche als Vergleichsbasis für geplante weitere Erhebungen in den nächsten Strategieperioden konzentriert.

3 Lernerfahrungen und Ansätze für eine Weiterentwicklung des GDA-Evaluationskonzepts für künftige Strategieperioden

Im Laufe der 1. GDA-Periode wuchs im »Steuerungskreis Dachevaluation«, dem zentralen Entscheidungsgremium für Fragen der Gesamtevaluation, in dem sowohl die drei GDA-Träger, als auch die Spitzenverbände von Arbeitgeber- und Arbeitnehmervertretungen mitarbeiten, das Bewusstsein für das folgende Defizit eines vorrangig auf die Steuerung durch Ziele orientierten Evaluationsansatzes: Das in der Evaluationsforschung beschriebene »klassische Zielmodell«, nach dem »der Grad der tatsächlichen Zielerreichung nur auf den beabsichtigten Zieldimensionen mit Hilfe eines Soll-Ist-Vergleichs bestimmt (wird) (Stockmann 2006, S. 179)«, birgt demnach »die Gefahr, dass nicht-intendierte Effekte systematisch ausgeblendet werden. Doch gerade diese können sich als ausgesprochen interessant und wichtig erweisen« (a. a. O., S. 180). In der ersten GDA-Periode wurde die Beobachtung dieser nicht-intendierten Effekte zwar nicht bewusst ausgeschlossen. Mit den fünf Zielebenen des ToR-Papiers als zentralem Steuerungsinstrument der GDA-Dachevaluation liegt bisher aber auch kein ausgearbeiteter Evaluationsleitfaden vor, der die Erhebung nicht-intendierter Effekte an relevanten Punkten und in strukturierter Weise erlauben würde.

Darüber hinaus wird immer klarer, dass wesentliche Voraussetzungen des klassischen Zielmodells v. a. im Bereich der Programmtheorie und der Definition bzw. der einheitlichen Operationalisierung von programmübergreifend festgelegten Indikatoren in der ersten Strategieperiode nicht umfassend gewährleistet werden konnten. Statt eine einheitliche Definition für die – als zentrale Indikatoren für das Ziel »MSE« verwendeten – Begriffe »Präventionskultur« und »Gesundheitskompetenz« vorzugeben, forderten die Arbeitsprogramme größtmögliche Souveränität für sich ein und wurden durch das ToR-Papier nur auf das weit gefasste Handlungsmodell nach Schweer, Krummreich (Schweer, Krummreich 2009) verwiesen. Die Mehrdimensionalität dieses auf Nutbeam Bezug nehmenden, an sich bestechenden Modells verleitete die beteiligten Akteure in insgesamt sechs dem »MSE«-Ziel zugeordneten Arbeitsprogrammgruppen zu sehr disparaten Operationalisierungsschwerpunkten. So knüpften die einen an bestimmte Begriffstraditionen in einzelnen Branchen an oder ließen sich vorrangig vom Gebot größtmöglicher Praxistauglichkeit und Niederschwelligkeit leiten, wodurch Aspekte der klassischen Arbeitsschutzorganisation mit Präventionskultur nahezu in eins gesetzt wurden. Andere Arbeitsprogrammgruppen wandten sich dem anspruchsvollsten Ende der Skala zu, d. h. der nach Nutbeam (Nutbeam 2000) höchsten Entwicklungsstufe »kritischer Gesundheitskompetenz« bzw. »kritischer Präventionskultur«, und fragten folglich direkt nach betrieblichen Arbeitsschutzmanagementsystemen, nach Mitarbeiterbefragungen und Dimensionen von Health Literacy bei Beschäftigten sowie nach betrieblicher Gesundheitsförderung (BGF). Sie nahmen diese fortgeschrittenen Aspekte in den Fokus ihrer Betriebsberatungen und folglich auch ihrer Evaluation.

Da die Operationalisierungen von Präventionskultur und Gesundheitskompetenz in den Arbeitsprogrammgruppen derart weit auseinander liegen, kann heute von einer Verwendung dieser Begriffe als »Indikatoren« für die MSE-Zielerreichung nur noch sehr bedingt die Rede sein. Allenfalls können die Items der in den Arbeitsprogrammen verwendeten MSE-Fachdatenbögen nach basalen Aspekten der Arbeitsschutzorganisation oder nach weiterführenden Aspekten von gesundheitsfördernder Organisationsentwicklung gruppiert und ggf. innerhalb der jeweiligen Gruppe vergleichend gegenüber gestellt werden, ohne dass allerdings abschließend geklärt wäre, welchen Beitrag etwaige branchenbezogene Einzelerfolge auf diesen Ebenen zur Entwicklung der Inzidenz von arbeitsbedingten Muskel-Skelett-Erkrankungen insgesamt leisten.

Dies ist eine generelle Schwachstelle der GDA und ihrer Dachevaluation. Eine bessere Ausarbeitung der jeweiligen Programmtheorien für künftige GDA-Perioden könnte diese Schwachstelle wenn schon nicht abschließend »heilen«, so doch durch begründete Hypothesen bzgl. anzunehmender Wirkungsketten besser ausfüllen.

4 Alternative: Theoretische Ansatzpunkte und die Leistungsfähigkeit eines stärker prozess- statt zielorientierten systemischen Evaluationsansatzes

Im Unterschied zur angestrebten Evidenzbasierung in der medizinischen Wirkungsforschung ist bei der Evaluation komplexer Interventionen im Bereich betrieblicher Prävention und Gesundheitsförderung a priori von einem Umstand auszugehen, der in der medizinischen Forschung zu einer deutlichen Herabstufung der Evidenz-Qualität führen würde (vgl. Schünemann 2009). Mit diesem Umstand ist die Eigenschaft komplexer Adaptivität einer heterogenen, aus sozialen Systemen (Betrieben) bestehenden Zielgruppe gemeint, die erwartbar zu heterogenen oder inkonsistenten Ergebnissen führt, ohne dass die Confounder und deren Wechselwirkungen untereinander abschließend bestimmt werden könnten. In einem solchen Feld liegt die Komplexität m. E. weniger in der Komplexität des Programms, das zumindest soweit es sich um die Einzelmaßnahme eines Arbeitsprogramms handelt, prinzipiell in überschaubarem Rahmen gehalten werden kann, als vielmehr in der nicht reduzierbaren Heteronomie von Betrieben als höchst unterschiedlichen sozialen Systemen, die sich fundamental von den psychophysischen Einheiten menschlicher Individuen unterscheiden. Letztere ähneln sich zumindest in den basalen Funktionsweisen ihres Stoffwechsels, während das Funktionieren eines Betriebes, die betriebsinterne Willensbildung sowie die im Rahmen des einzelnen Betriebes tatsächlich gelebte Führung und Zusammenarbeit sich abhängig von Betriebsgröße und Branche, Arbeitsaufgabe, Bildungsstand, Altersstruktur und Geschlecht der Mehrheit seiner Beschäftigten – um nur einige der möglichen Confounder zu nennen – so gravierend unterscheiden, dass inkonsistente Reaktionsweisen auf die in die Betriebe getragenen Stimuli nicht die Ausnahme sein können, sondern als die Regel betrachtet werden müssen.

Dazu schreibt Richard Hummelbrunner, der in Bezug auf komplexe Interventionen zu systemischen Evaluationsansätzen rät und dafür methodische Werkzeuge präsentiert: »Unter diesen Bedingungen können Ergebnisse von Programmen weder vorab festgelegt, noch auf die ursprüngliche Intention der Programme reduziert werden. Es ist schwierig, wenn nicht unmöglich, klare Kausalbeziehungen und Effekte zu beschreiben. Und genauso schwierig ist es, im Planungsstadium inhaltlich angemessene Ergebnis-Indikatoren und Zielniveaus zu bestimmen (Hummelbrunner 2011, S. 93; aus dem Englischen von B. S.).«

Die Evaluation von Programmen bzw. einer Strategie, die auf komplex adaptive Zielgruppen einwirkt, ausschließlich durch Impact-Indikatoren zu steuern, ist deshalb für Hummelbrunner »weder machbar noch angemessen (a. a. O.)«. Ein geeignetes Monitoring-System müsse vielmehr Informationen generieren, die den Programm-Akteuren hilft, den Status der Faktoren zu erkennen, die direkt von ihnen beeinflusst werden können und zugleich als entscheidende Schlüsselfaktoren für die Zielerreichung anerkannt sind. Allgemein formuliert betreffen diese Schlüsselfaktoren immer die Qualität (das Wie?) der Implementierungs-Aktivitäten, das organisatorische Prozedere eines Programms sowie die tatsächlich beobachteten Verhaltensänderungen von Partnern und Zielgruppen.

Kurz gesagt: Evaluationen von komplexen Interventionen, die auf komplex adaptive soziale Systeme einwirken, müssen im Sinne eines systematischen Monitorings mehr die – tatsächlich stattfindenden – Prozesse, denn die deklamatorischen Ziele und deren Indikatoren betrachten. Bei den zunächst theoretisch zu identifizierenden und dann empirisch bzgl. ihres tatsächlichen Stattfindens zu verifizierenden Prozessen muss es sich um solche Prozesse handeln, »von denen erwartet wird, dass sie zu Ergebnissen bzw. Wirkungen (Impacts) führen (a. a. O.)«. Eine Evaluation, die sich auf diese Prozesse konzentriert, kann und darf sich nicht darauf beschränken, am Ende Indikatoren für die Zielerreichung zu messen. Zumal die Information, die eine solche Indikatoren-Messung generiert, in aller Regel zu spät kommt und nichts darüber aussagt, wie die für den Programmerfolg entscheidenden Prozesse ggf. hätten verändert werden müssen, um einen höheren Zielerreichungsgrad zu ermöglichen.

Wenn dieser prozessorientierte Ansatz im Falle der GDA ergänzend zu einem modifizierten ToR-Papier in der nächsten GDA-Periode ab 2013 zum Zuge kommen soll – und die letzten Beschlüsse der Nationalen Arbeitsschutzkonferenz (NAK) weisen in diese Richtung –, heißt dies, es müssten bzgl. der von Hummelbrunner genannten generellen

Erfolgsfaktoren Implementierung, Organisation und Zielgruppenverhalten u. a. folgende Evaluationsfragen gestellt werden:

- **Implementierung:** Wie werden die GDA-Arbeitsprogramme umgesetzt? Wie genau handhaben die Aufsichtsdienste die Leitlinien?
- **Organisation:** Wie funktionieren die organisatorischen Rahmenbedingungen der GDA? Wie funktioniert beispielsweise die vertikale Kommunikation über GDA-Leitlinien von den Amts- bzw. Präventionsleitungen zu den Aufsichtsdiensten vor Ort?
- **Zielgruppenverhalten:** Welche Verhaltensmuster dominieren bei der Umsetzung von Arbeitsschutzanforderungen in Klein- bzw. Großbetrieben? Welche Überwachungsschritte oder Präventionsangebote werden von welchen Betrieben stärker genutzt bzw. führen stärker zu einer Verhaltensänderung als andere?

Die Liste dieser Fragen kann prinzipiell beliebig fortgesetzt werden. Unter allen denkbaren Fragen zu den o. g. Erfolgsfaktoren ist aber eine sinnvolle Auswahl zu treffen, die sich ergibt, wenn man folgenden methodischen Schritten folgt:

a) Identifizieren und Strukturieren von intendierten Effekten,
b) Ausformulierung von Hypothesen bzgl. des Erreichens der Effekte,
c) Festlegen von zu beobachtenden Bereichen für das Monitoring von Schlüsselprozessen und
d) Sammeln und Interpretieren von Daten.

Zu diesen methodischen Schritten wird derzeit von einer Unterarbeitsgruppe des Steuerungskreises GDA-Dachevaluation ein Evaluations-Leitfaden erarbeitet, der die Liste der o. g. Evaluationsfragen fortsetzt und eine Anleitung zur Bildung entsprechender Wirkungshypothesen sowie zur Identifizierung von Schlüsselprozessen und zur Definition von Indikatoren liefern wird. Der geplante Evaluations-Leitfaden ist als Leitfaden für die Gesamtevaluation der GDA in der 2. Periode angedacht. Er soll aber auch auf die Arbeitsprogramme anwendbar sein, was der Absicht der GDA-Träger entspricht, die Arbeitsprogramm-Evaluationen künftig besser mit der GDA-Dachevaluation zu verzahnen. Die Konzepte der Arbeitsprogramm-Evaluationen sowie der Gesamtevaluation in der zweiten GDA-Periode werden anhand des künftigen Leitfadens zu präzisieren sein, wobei bezüglich der zu perpetuierenden Repräsentativbefragungen von Betrieben und Beschäftigten (Bestandteil der Gesamtevaluation) darauf zu achten ist, dass die Vergleichbarkeit zwischen Erst- und Zweiterhebung gewahrt bleibt.

5 Weiterführende Überlegungen

Es dürfte bereits deutlich geworden sein, wie wichtig die genaue Beobachtung bzw. das genaue Monitoring der Schlüsselprozesse ist, die die GDA-Träger und Sozialpartner – gemeinsam in einem partizipativen Prozess – hypothetisch als wesentliche Wirkungspfade (Schlüsselprozesse) identifizieren werden. Das genaue Beobachten dieser Prozesse dient einerseits der Hypothesenprüfung, andererseits eröffnet es die Möglichkeit, Ressourcen ggf. noch während der Strategieperiode umzusteuern, sofern sich andere als die angenommenen oder nur eine Teilmenge der angenommenen Schlüsselprozesse als besonders wirksam im Sinne des aktiven Aufgreifens der Interventionen durch die Betriebe zeigen sollten.

Bei komplexen Interventionen in der Prävention geht es – das zeigt die bisherige Erfahrung mit der Evaluation der GDA – m. E. vorrangig um den Versuch, erwünschte innerbetriebliche Prozesse »von außen« zu initiieren. Dabei fungieren die Betriebe in ihrer Eigenschaft als komplex adaptive soziale Systeme als Ko-Produzenten der angestrebten Wirkungen (Impacts). Mit der Souveränität und Eigenwilligkeit der Betriebsleitungen ist also von vornherein zu rechnen. Der Subjektstatus muss ihnen deshalb im Rahmen des hier vertretenen systemischen Evaluationsansatzes sowohl theoretisch, als auch in der Praxis zugebilligt werden, was einer zunächst denkbaren, wenn auch forschungspraktisch äußerst schwer umzusetzenden Randomisierung im Rahmen eines Kontrollgruppenansatzes m. E. enge Grenzen setzt – nicht zuletzt auch, weil durch bestimmte Formen der Randomisierung die Erzielung der größtmöglichen Wirkung in Frage steht.

Ein kurzer Exkurs zu einem aktuellen Versuch, in der betrieblichen Präventionsforschung mit einem Kontrollgruppendesign zu arbeiten, soll dies verdeutlichen: im Rahmen der Kreuzinterventionsstudie »KRISTA« der BGW wurde Altenpflegeheimen mit der Zuordnung zur Kontroll- oder zur Interventionsgruppe durch ein Zufallsverfahren jeweils eine Rückenschule für die Beschäftigten oder eine Intervention zum Hautschutz in der Altenpflege zugeordnet. Der Widerstand von Heimen, die zufällig für den Hautschutz ausgewählt wurden, obwohl ihre Motivation zur Teilnahme ursprünglich aus dem Wunsch nach einer Rückenschule erwuchs – und vice versa –, wurde nur zum Teil überwunden. Dort, wo Betriebe dennoch zu einer Teilnahme bewegt werden konnten, gelang dies »nur durch die Ankündigung einer Wiederholung der beiden Präventionsmaßnahmen nach Abschluss der Studie (Dulon 2011, S. 87).« Hier muss m. E. gefragt werden, ob das gewählte Forschungsdesign, das im Kontext medizinischer Wirkungsforschung seine Berechtigung haben mag, im Feld der betrieblichen Präventionsforschung nicht Gefahr läuft, selbst den Bias zu produzieren, den es zu vermeiden sucht. Damit meine ich den – die Wirkung potenziell abschwächenden – Effekt mangelnder Motivation mehr oder weniger »zu ihrem Glück« gezwungener Betriebsleitungen und ihrer Beschäftigten. Sobald man aufhört, die Verzerrungen durch diesen und ähnliche negative Effekte von Randomisierungen zu unterschätzen, kommt man nicht umhin die Frage zu stellen, ob es im Bereich der Evaluation »von außen« initiierter betrieblicher Präventionsfortschritte nicht fahrlässig sein kann, zugunsten des vermeintlich besten Forschungsdesigns auf die Ko-Produktion der Ergebnisse (Outcomes) und mittelbar auch der Impacts durch interessierte und motivierte Betriebe zu verzichten. Die größtmögliche Wirkung zumindest dürfte man mit einem solchen Vorgehen nicht erzielen, was durchaus mit dem Umstand in Verbindung gebracht werden kann, dass die Ergebnisse von Studien, die dem vermeintlichen »wissenschaftlichen Goldstandard« folgen, im Bereich der betrieblichen Präventionsforschung oft genug eben nicht die Evidenz-Qualität produzieren, die ihrem eigenen Anspruch entspricht.

Mit Blick auf die GDA erschöpfen sich die Vorbehalte gegenüber einem randomisierten, kontrollierten Forschungsdesign jedoch nicht in diesen Überlegungen. Hinzu kommt vielmehr, dass die GDA in den Betrieben umgesetzt wird »durch ›Verwaltungshandeln‹ in der Anwendung von bestehendem Recht. Es kann deshalb nicht eine Gruppe von Betrieben und Arbeitnehmern ausgeschlossen werden, um eine ›Kontrollgruppe‹ zu erhalten. Das verbietet das Gebot der Gleichbehandlung bei der Rechtsanwendung (Stamm et al. 2011, S. 444)«.

Jeder Versuch einer Randomisierung in dem in Frage stehenden Forschungsfeld muss diese beiden Rahmenbedingungen berücksichtigen: Betriebe dürfen durch die Randomisierung weder demotiviert noch komplett von Maßnahmen der Aufsichtsdienste ausgeschlossen werden. Dennoch kann auch unter diesen Randbedingungen in engen Grenzen mit dem Kontrollgruppenansatz gearbeitet werden. Das zeigt eine wissenschaftlich begleitete Schwerpunktaktion der österreichischen Arbeitsschutzverwaltung. Diese hat beispielsweise eine Gruppe von Betrieben vor Ort besucht und zum Explosionsschutz beraten, eine zweite hingegen nur per E-Mail über die Erfordernisse des Explosionsschutzes informiert. Anschließend wurde eine Zweiterhebung in beiden Gruppen sowie eine Erhebung in einer dritten Gruppe von Betrieben vorgenommen, die weder persönlich beraten noch per E-Mail informiert wurden (vgl. Kerschhagl et al. 2010). Ziel der Untersuchung war, die stärkere Wirksamkeit der persönlichen Beratung vor Ort nachzuweisen, was auch gelungen ist. Mit diesem anders fokussierten Vorgehen wird das soziale System Betrieb als eine mündige Organisation und als Ko-Produzent der angestrebten Impacts ernst genommen und keine Gruppe von Betrieben systematisch ausgeschlossen. Beides ist funktional für den größtmöglichen Erfolg der Maßnahme – und damit auch ihrer Evaluation.

Auf die Anwendung von Kontrollgruppendesigns – gerade bei der Evaluation von Instrumenten, Ansätzen, Prozessen – muss also nicht gänzlich verzichtet werden. Bei erfolgreicher Anwendung, d. h. wenn tatsächlich ein deutlicher Wirkungsunterschied zwischen Interventions- und Kontrollgruppe nachgewiesen werden kann, ist aufgrund der multiplen Wirkungsketten innerhalb komplex adaptiver sozialer Systeme allerdings noch längst nicht klar, was genau zur besseren Bilanz der Interventionsgruppe beigetragen hat. Diese sogenannte black box kann durch eine – im Bereich betrieblicher Präventionsforschung ohnehin nur begrenzt

mögliche – Confounder-Kontrolle nur verkleinert, aber nicht abschließend beseitigt werden.

Auch deshalb sollte die betriebliche Präventionsforschung m. E. vorrangig mit systemischen Methoden ihre eigene Evidenz generieren. Systemische Evaluationsmethoden sollten im Vordergrund stehen, wiewohl sie für einige klar abgegrenzte, d. h. weniger komplexe Maßnahmen durch Wirkungsmessungen mit dem Kontrollgruppenansatz ergänzt werden können. Das hier postulierte Primat systemischer Methoden leitet sich letztlich aus der Maxime ab, dass die Wahl der Methode dem Gegenstand folgen, d. h. diesem angemessen sein sollte. Das Prozessorientierte Wirkungsmonitoring (PWM) beispielsweise eignet sich nach Auskunft seiner Entwickler v. a. für die folgenden – komplexen – Evaluationsvorhaben:

- Interventionen mit langfristigen Wirkungsketten, die dadurch gekennzeichnet sind, dass die angestrebten Erfolge (impacts) erwartbar erst am Ende der Implementierungsphase oder noch später auftreten,
- »weiche« und »open-ended« Interventionen (z. B. Förderung von Innovation oder Wettbewerbsfähigkeit),
- Impact-orientierte Förderprogramme, die rechtzeitig Informationen über den voraussichtlichen Erfolg von Maßnahmen und Projekten für ihre Steuerung benötigen und
- Programme, die durch eine große Zahl von unabhängigen Akteuren und Projekten umgesetzt werden. Hypothesen über die für den Erfolg entscheidenden Schlüsselprozesse dienen hier als gemeinsame Regeln, um bezüglich der angestrebten Effekte Kurs zu halten (vgl. Hummelbrunner 2011, S. 104).

Das PWM oder auch andere systemische Methoden ermöglichen die Evaluation eines Bündels von hypothetisch aufgestellten und dann empirisch verifizierten Wirkungsketten, an deren Beginn der Einfluss der Interventionen der GDA noch relativ unmittelbar verfolgt und nachgewiesen werden kann, an deren Ende jedoch mittelbare Wirkungen stehen. Dabei geht es der Methode im Wesentlichen um die Identifizierung von Prozessen, die relevant sind für die Erzielung von Resultaten bzw. Impacts, und um ein Monitoring, das klärt, inwieweit die Prozesse stattgefunden haben und ob deren Relevanz für die Zielerreichung tatsächlich gültig ist oder ob nicht-intendierte Effekte ggf. eine größere Rolle spielen. Dies ermöglicht entsprechende Rückkopplungsschleifen im Strategieverlauf und legt den Schwerpunkt auf die Prozessevaluation, von der ausgehend mit Plausibilitätsschlüssen auf die Wirkung geschlossen wird. Mit den Ergebnissen einer solchen prozessorientierten Wirkungsforschung kann m. E. die black box des Kontrollgruppenansatzes in der betrieblichen Präventionsforschung ausgeleuchtet werden, so dass beide Ansätze sich auch in diesem Sinne gegenseitig ergänzen können.

Nachfolgend sind die beiden diskutierten Ansätze der Wirkungsmessung in der Evaluation noch einmal grafisch gegenüber gestellt.

Abbildung 1
Kontrollgruppendesign

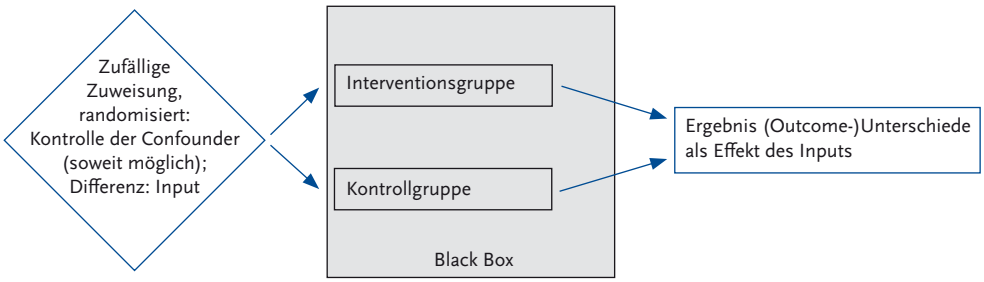


Abbildung 2
Prozessorientierte systemische Evaluation

PWM-Wirkungs-Modell →	Leistung →	Output →	Nutzung (kurzfristig) →	Ergebnis (mittelfristig) →	Wirkung
Beschreibung	Aktivitäten und Leistungen der GDA-Akteure (Träger und Kooperationspartner)	Transfer der Leistungen der GDA-Akteure in die Betriebe/ Unternehmen	erwartetes Aufgreifen der Leistungen der GDA-Akteure durch die Betriebe/ Unternehmen (betriebliche Entscheidungen)	beabsichtigte betriebliche Effekte (Outcomes): Wurden betriebliche Maßnahmen ergriffen in Form von Aktionen, Anschaffungen, Änderungen der Aufbau- oder Ablauforganisation?	längerfristig beabsichtigte Wirkungen (Impacts) in Betrieben und in der Gesellschaft
	Ausleuchten der Black Box: Beobachtung der Mikroprozesse der Intervention		Plausiblitätsschluss: Wenn diese Nutzung durch die Betriebe erzielt wird, wird es positiv in Richtung Ergebnis wirken (Ko-Produktion des Ergebnisses)		

Literatur

Dulon M, Wendeler D, Nienhaus A (2011) Sind randomisierte kontrollierte Studien in der Altenpflege möglich? – Erfahrungen aus der KRISTA-Studie. Zbl Arbeitsmed 61: 84–87

Gemeinsame Deutsche Arbeitsschutzstrategie (Hrsg) (2007) Fachkonzept und Arbeitsschutzziele 2008 – 2012, Berlin www.gda-portal.de (Stand: 12.12.2007)

Hummelbrunner R (2011) Process Monitoring of Impacts. In: Williams B, Hummelbrunner R: Systems Concepts in Action – A Practioner's Toolkit. Stanford University Press, Stanford, S 92–107

Kerschhagl J, Neuwirth E, Jauering P et al., Bundesministerium für Arbeit, Soziales und Konsumentenschutz (Hrsg) (2010) Explosionsschutz in kleineren und mittleren Unternehmen. Realisierung des Explosionsschutzes – Wirkung von Beratung und Information, eine Schwerpunktaktion im Rahmen der Arbeitsschutzstrategie, durchgeführt in den Branchen KFZ-Lackierereien und Tischlereien. Wien www.arbeitsinspektion.gv.at (Stand: Dezember 2010)

Lißner L, Reihlen A, Höcker H et al. (2010) Vergleichende Analyse nationaler Strategien für Sicherheit und Gesundheit bei der Arbeit, BAuA-Forschungsbericht 2234, Dortmund/Berlin/Dresden

Nutbeam D (2000) Health Literacy as a public health goal: a challenge for contemporary health education and communication strategies into the 21st century. Health Promotion International 15 (3): 259–267

Schünemann H (2009) GRADE: Von der Evidenz zur Empfehlung – Beschreibung des Systems und Lösungsbeitrag zur Übertragbarkeit von Studienergebnissen. Z. Evid. Fortbild. Qual. Gesundh. wesen (ZEFQ) 103 (6): 391–400

Schweer R, Krummreich U (2009) Gesundheitskompetenz und Präventionskultur – Indikatoren für Gesundheit und Erfolg in Unternehmen: ein praktisches Handungsmodell. Z.Arb.Wiss. 63: 293–302

Stamm R, Lenhardt U, Pernack E et al. (2011) Die Evaluation der Gemeinsamen Deutschen Arbeitsschutzstrategie – GDA. Sicher ist Sicher – Arbeitsschutz aktuell 10: 442–447

Stockmann R (2006) Evaluation und Qualitätsentwicklung. Eine Grundlage für wirkungsorientiertes Qualitätsmanagement, Sozialwissenschaftliche Evaluationsforschung Bd 5, Waxmann, Münster

Evaluation komplexer Interventionsprogramme in der Prävention: Das Beispiel IN FORM

Ute Winkler, Barbara Werner-Schlechter, Diana Hart

Allgemein gilt, dass nur ein lernendes System, welches seine Prozesse regelmäßig überprüft, damit seine Stärken und Schwächen kennt und sich ständig weiter entwickelt, den Ansprüchen an gute Prävention zum Wohle der Bürgerinnen und Bürger gerecht werden kann. Am Beispiel des Nationalen Aktionsplans »IN FORM – Deutschlands Initiative für gesunde Ernährung und mehr Bewegung« sollen die Möglichkeiten, aber auch Grenzen der Evaluation aufgezeigt werden.

Der Nationale Aktionsplan »IN FORM – Deutschlands Initiative für gesunde Ernährung und mehr Bewegung« wurde am 25. Juni 2008 vom Bundeskabinett verabschiedet. Er verfolgt das Ziel bis 2020, das Ernährungs- und Bewegungsverhalten in Deutschland nachhaltig zu verbessern. Er wird in gemeinsamer Federführung von Bundesministerium für Gesundheit (BMG) und Bundesministerium für Ernährung, Landwirtschaft und Verbraucherschutz (BMELV) durchgeführt.

Konkret heißt es im Aktionsplan: Es soll erreicht werden, dass Erwachsene gesünder leben, Kinder gesünder aufwachsen und von einer höheren Lebensqualität und einer gesteigerten Leistungsfähigkeit in Bildung, Beruf und Privatleben profitieren und Krankheiten deutlich zurückgehen, die durch einen ungesunden Lebensstil mit einseitiger Ernährung und Bewegungsmangel mit verursacht werden. Bis zum Jahre 2020 sollen sichtbare Ergebnisse erreicht werden. Gesundheitsberichte sowie die Ergebnisse regelmäßiger Monitorings sind wichtige Instrumente, um den Erfolg zu dokumentieren (vgl. www.in-form.de).

Das sind die selbst gesetzten Ziele, an denen sich die Aktivitäten und Erfolge bis 2020 messen lassen müssen.

Um die Komplexität des Aktionsplans und die damit verbundenen Herausforderungen und Schwierigkeiten im Hinblick auf eine Evaluation zu verdeutlichen, sollen zunächst kurz die Aktivitäten und Arbeitsstrukturen beschrieben werden.

Von Beginn an war es die Intention von IN FORM, die parallel existierenden vielfältige Aktivitäten von Bund, Ländern, Kommunen und Zivilgesellschaft in einer nationalen Strategie zu Stärkung und Etablierung gesundheitsförderlicher Alltagsstrukturen in den Bereichen Ernährung und Bewegung zusammenzuführen und fortzuentwickeln.

Der Nationale Aktionsplan ist ein langfristig angelegter Prozess und ein dynamisches Instrument des Dialoges, der mit allen relevanten Akteuren umgesetzt und weiterentwickelt werden muss. Dazu wurden Strukturen zur Einbindung aller politischen Ebenen und relevanter Politikbereiche sowie der Zivilgesellschaft, Wirtschaft und Wissenschaft geschaffen. So wurde z. B. auf Bundesebene eine interministerielle Arbeitsgruppe geschaffen, um eine engere Kooperation der einzelnen Ressorts sicherzustellen. Da auch die Länder und Kommunen eine Vielzahl an Aktivitäten ergriffen haben und teilweise auf Grund des föderalen Systems die Zuständigkeit z. B. für den Bereich Schule haben, werden diese ebenfalls in einer Bund-Länder-Arbeitsgruppe in den Prozess einbezogen. Weitere wichtige gesellschaftliche Akteure wie die Sozialversicherungsträger, die Sportverbände, die Lebensmittelindustrie, Verbraucherverbände, Arbeitgeber- und Arbeitnehmerverbände, medizinische Fachgesellschaften u. ä. sind in der Nationalen Steuerungsgruppe vertreten, der die inhaltliche Impulsgebung und die Vernetzung der relevanten gesellschaftlichen Akteure obliegt. Eine Geschäftsstelle koordiniert die Aktivitäten und unterstützt die Ministerien. Zur Unterstützung von Vernetzungsaktivitäten und der Verbreitung von Informationen zum Prozess wurde ein Internet-Portal erstellt, das sich getrennt einerseits an »Profis« – also an Multiplikatoren und mögliche Akteure vor Ort wendet und andererseits die Informationen für die Bürgerin/den Bürger zur Verfügung stellt.

Zu einzelnen Themenbereichen des Aktionsplans wurden Arbeitsgruppen eingerichtet, die Vorschläge zur inhaltlichen Gestaltung erarbeiten. So haben Verbände und Expertinnen und Experten Empfehlungen zur Umsetzung von Bewegungsförderung im Alltag erstellt.

Die beiden Ministerien haben seit 2008 rund 100 Projekte im Rahmen von IN FORM finanzi-

ell unterstützt. Aufbauend auf den Zielsetzungen des Aktionsplans sind in einem Zusammenspiel von Verhaltens- und Verhältnisprävention einerseits Maßnahmen zur Stärkung und Etablierung gesundheitsförderlicher Alltagsstrukturen initiiert worden und andererseits Aktivitäten unternommen worden, um die Kenntnisse über die Zusammenhänge von Ernährung, Bewegung und Gesundheit zu verbessern.

So wurden von 2008 bis 2011 seitens des BMG elf »Aktionsbündnisse gesunde Lebensstile und Lebenswelten« ins Leben gerufen, die bestehende Angebote vor Ort vernetzen und unterschiedliche Akteure auf regionaler bzw. lokaler Ebene zusammenbringen. Um Alltagsbewegung möglichst breit zu verankern, sind in allen Ländern »Zentren für Bewegungsförderung« eingerichtet worden. Diese geben einen Überblick über die Angebote, informieren über gute Praxisbeispiele und stehen als kompetente Ansprechpartner zur Verfügung. Vorrangig sollen hier ältere Menschen angesprochen werden.

In Modellprojekten wurden besondere Aspekte zur Herstellung von gesundheitlicher Chancengleichheit berücksichtigt. Es wurden einzelne Zielgruppen wie Kinder, Ältere, sozial Benachteiligte und Menschen mit Migrationshintergrund angesprochen, übergeordnete Fragestellungen z. B. zur Qualitätssicherung bearbeitet und konkrete Veränderungen in den Lebenswelten wie Kindertagesstätten, Schulen, Betrieben und Senioreneinrichtungen aufgegriffen.

Vor dem Hintergrund dieser komplexen Aktivitäten und Strukturen und der erklärten Zielsetzung bis 2020 stellt sich nun die Frage, in welcher Form eine Evaluation in diesem Prozess denkbar und möglich ist. Zunächst soll 2014 ein Zwischenbericht vorgelegt werden, anhand dessen eine erste Bilanz gezogen werden soll.

Eine seriöse wissenschaftliche Evaluation des Gesamtprozesses ist auf Grund der Vielzahl der verschiedenen unmittelbaren und mittelbaren Einflüsse kaum möglich. In den genannten sozialen komplexen Systemen gibt es eine hohe Dynamik und viele externe Einflussfaktoren, so dass Entwicklungen nicht vorhersehbar sind. So lassen sich Evaluationsergebnisse aus einzelnen Projekten nicht einfach generalisieren und auf andere Kontexte übertragen. Es erscheint kaum möglich, festgestellt Effekte einzelnen Interventionen zuzuschreiben. Dies bedeutet jedoch nicht, dass die Maßnahmen von IN FORM nicht in der fachlich gebotenen Qualität und nach dem allgemein anerkannten wissenschaftlichen Stand angeboten werden müssen.

Die Qualitätssicherung des IN FORM-Prozesses wird auf drei Ebenen erbracht.

- Ein Großteil der Teil der IN FORM-Projekte wurde von der Art und vom Umfang her unterschiedlich evaluiert. Insbesondere die elf »Aktionsbündnisse für gesunde Lebenswelten und Lebensstile« wurden während der Projektlaufzeit von einem externen »Evaluationsprojekt« bündnisübergreifend betreut. Hierzu liegt ein Abschlussbericht aus dem Jahr 2011 vor, der auf den Internetseiten des BMG bzw. IN FORM abrufbar ist.
- Auf Projektebene bietet IN FORM außerdem Hilfestellung an, eine bessere Qualitätssicherung und Evaluation zu ermöglichen. Dies ist das erklärte Ziel der Arbeitsgruppe »Qualitätssicherung«. Die Veröffentlichung der Toolbox »Projekte IN FORM – Wege zur Qualität« und die Umsetzung entsprechender Anforderungen bei neuen Projekten sind wichtige Schritte in diese Richtung. Diese Toolbox will einen niedrigschwelligen Zugang zum Thema Qualitätssicherung sowohl für kleine wie auch große Praxisprojekte anbieten. Die Arbeitsgruppe hat sich zum Ziel gesetzt, eine entsprechende Toolbox zum Thema Evaluation zu entwickeln und für Projekte zur Verfügung zu stellen.
- Schließlich wird auf Metaebene ein Set von Kriterien und Indikatoren benötigt, das geeignet ist, die Veränderungen des Ernährungs- und Bewegungsverhaltens in Deutschland abzubilden. Im Rahmen der Gesundheitsberichterstattung des RKI (DEGS, KiGGS und Folgewellen, GEDA) und der Nationalen Verzehrstudie sowie des Ernährungsmonitorings des MRI liegen entsprechende Kriterien und Indikatoren vor, anhand derer entsprechende Veränderungen verfolgt und beobachtet werden können. Dazu gehört der Body Mass Index (BMI), die körperliche Aktivität, das Bewegungsverhalten (Anteil sportlicher Aktiver, körperliche Aktivität im Alltag), das Ernährungsverhalten (Häufigkeit und Menge des Obst- und Gemüsekonsums, Verzehr einer einzelnen Lebensmittelgruppe in g/Tag,

Index für gesundheitliche Qualität der Ernährung) und das eigene Gesundheitsempfinden (subjektive Gesundheit).

Klar ist auch hier, dass zukünftige Veränderungen innerhalb dieser Indikatoren nicht kausal auf Maßnahmen des IN FORM-Prozesses zurückzuführen sind und auch Rückschlüsse auf einzelne Interventionen nicht möglich sind, die Indikatorenauswahl und die Kontinuität ihrer Erhebung ermöglichen aber eine prozessbegleitende Verlaufsbeobachtung gesundheitsbezogener Verhaltensweisen im Sinne eines Hintergrundmonitorings auf Bevölkerungsebene.

Evaluation der Gesundheitsinitiative Gesund.Leben.Bayern.

Veronika Reisig

1 Einführung

Die Gesundheitsinitiative Gesund.Leben.Bayern. (GLB) wurde im September 2004 durch einen Ministerratsbeschluss ins Leben gerufen. Hauptanliegen war, die Prävention und Gesundheitsförderung in Bayern effektiver zu gestalten, u. a. durch eine stärkere Konzentration auf prioritäre Handlungsfelder und Zielgruppen. In den Konzeptentwicklungsprozess der Initiative flossen die Ergebnisse verschiedener Arbeitsstränge ein, u. a. die Ergebnisse eines 2004 durchgeführten Bürgergutachtens zur Gesundheit (Sturm, Weilmeier 2004), einer ebenso 2004 ausgeführten Erhebung der Präventionsaktivitäten in Bayern (Wildner et al. 2006) sowie die Aufarbeitung epidemiologischer Daten zu den wichtigsten Gesundheitsrisiken in Bayern. Ein darauf aufbauendes, in Expertenworkshops entwickeltes und durch gesellschaftliche und politische Akteure konsentiertes Grundlagenpapier legt den konzeptionellen und inhaltlichen Rahmen der Gesundheitsinitiative fest. Konsentierte Leitprinzipien der Initiative sind (Wildner et al. 2007):

- Konzentration auf prioritäre Handlungsfelder und Zielgruppen,
- Fokus auf Evidenzbasierung, Evaluation und Qualität,
- Priorität von Maßnahmen der Gesundheitsförderung und Primärprävention (gegenüber sekundär- und tertiärpräventiven Ansätzen),
- Bevorzugung von lebensweltorientierten Maßnahmen,
- Ausweitung bewährter Maßnahmen,
- Begleitung der Initiative durch die Gesundheitsberichterstattung.

Als Handlungsebenen der Gesundheitsinitiative sind die Förderung von Modellprojekten, Öffentlichkeitsarbeit und Agendasetting sowie Vereinbarungen und Rechtsetzung vorgesehen.

Die Schwerpunkthandlungsfelder der Initiative waren anfänglich die Bereiche Rauchfrei Leben, Verantwortungsvoller Umgang mit Alkohol, Gesunde Ernährung und Bewegung sowie Gesunde Arbeitswelt. Diese wurden in den Jahren 2007 bzw. 2008 um die Handlungsfelder Gesundheit im Alter und Psychische Gesundheit erweitert. Die Handlungsfelder wurden mit mehr oder weniger expliziten Gesundheitszielen ausgestattet, die zum Teil (semi)quantitativ, zum Teil qualitativ formuliert sind (Wildner et al. 2007). Die Ziele für den Bereich Rauchfrei Leben zum Beispiel sind eine Trendwende beim Einstiegsalter bis 2008, Senkung der Jugendlichenraucherquote um 25 % bis 2015 sowie weitere Einzelziele zur Schaffung rauchfreier Lebenswelten in verschiedenen Settings. Darüber hinaus verfolgt die Initiative folgende nebengeordnete Zielsetzungen:

- einen Beitrag zur Evidenzbasis in der Gesundheitsförderung und Prävention (GFP) zu leisten,
- zur Qualitätsentwicklung der geförderten Projekte und zur Qualitätskultur im Feld der GFP beizutragen
- und Modelle guter Praxis zu identifizieren.

Mit ihrer strategischen Zielsetzung, den mehrfachen Handlungsebenen und insbesondere der Projektförderung, die eine Einbeziehung verschiedenster Akteure und Ansätze im Rahmen der verschiedenen Schwerpunkthandlungsfelder und der gemeinsamen strategischen Zielsetzungen impliziert, stellt die Initiative Gesund.Leben.Bayern. eine komplexe Intervention im Bereich der Prävention und Gesundheitsförderung dar.

Entsprechend den Leitprinzipien der Initiative ist die Evaluation der einzelnen Projekte eine Voraussetzung der Projektförderung. Darüber hinaus wurde auch eine begleitende externe Evaluation der gesamten Initiative mit all ihren Handlungsebenen im Zeitraum 2005 bis 2008 durchgeführt. Dieser Evaluationsansatz soll im Weiteren näher vorgestellt und diskutiert werden.

2 Evaluationsansatz und Methoden

Zwei-Ebenen-Evaluationskonzept

Gleichwertige Evaluationsziele waren zum einen die Qualitätsentwicklung bzw. -sicherung der geförderten Einzelprojekte sowie der Gesamtinitiative und zum anderen die summative Bewertung von Einzelprojekten und Gesamtinitiative im Hinblick auf die jeweils gesetzten Ziele. Dies bedeutet, dass die Evaluation auf zwei Ebenen, nämlich auf der der Einzelprojekte wie auch der der Gesamtinitiative stattfindet, wobei diese beiden Evaluationsebenen mehrfach miteinander verknüpft sind (Reisig et al. 2007). Die Gesamtinitiative wird hierbei als tragendes strategisches Grundgerüst für die Einzelprojektförderung verstanden.

Einzelprojektebene

Die durch Gesund.Leben.Bayern. geförderten Einzelprojekte sind von Anfang an in einen Qualitätsmanagementprozess eingebunden, beginnend mit der Projektbegutachtung (Planungsevaluation) im Rahmen der Antragstellung und gefolgt von der Bewertung der Zwischenberichte laufender Projekte (formative Evaluation). Bei Bedarf werden nach Begutachtung des Projektantrags bzw. des Zwischenberichts Empfehlungen an die Antragsteller bzw. Projektnehmer übermittelt. Zum Projektende wird eine abschließende, d. h. summative, Prozess- und Ergebnisevaluation erstellt. Diese wird von den Projektnehmern in Form einer Selbstevaluation durchgeführt, ggf. mit weiterer wissenschaftlicher Unterstützung. Die Ergebnisevaluation der Einzelprojekte soll nach Möglichkeit eine Vorher-Nachher-Untersuchung von Zielindikatoren in einem Kontrollgruppendesign beinhalten. Mittel für die Evaluationsaufgaben werden den Projektnehmern in angemessener Höhe mitbewilligt. Auf Grundlage der Selbstevaluation werden die Prozess-, Ergebnis- und Evaluationsqualität der Einzelprojekte mit einem eigens hierfür entwickelten, standardisierten Instrumentarium (Loss et al. 2007) extern bewertet (Fremdevaluation). Darüber hinaus wird für jedes geförderte Projekt betrachtet, inwiefern es einen Beitrag leistet zu den Gesundheitszielen der Initiative, zur Evidenzbasis in der GFP, zur Qualitätskultur und ob es als Modell guter Praxis weiterempfohlen werden kann. Diese Fremdevaluation ist Bestandteil der begleitenden externen Evaluation der Initiative und stellt einen Verknüpfungspunkt der beiden Evaluationsebenen »Einzelprojekte« und »Gesamtinitiative« dar. Einen Überblick über den Einzelprojektförderzyklus mit den verschiedenen Evaluationsschritten gibt Abbildung 1.

Ebene der Gesamtinitiative

Die Evaluation der Gesamtinitiative bestand aus einer externen Bewertung der Konzept-, Struktur-, Prozess- und Ergebnisqualität der Initiative nach übergeordneten Gesichtspunkten wie in Tabelle 1 dargestellt.

Abbildung 1
Einzelprojektzyklus

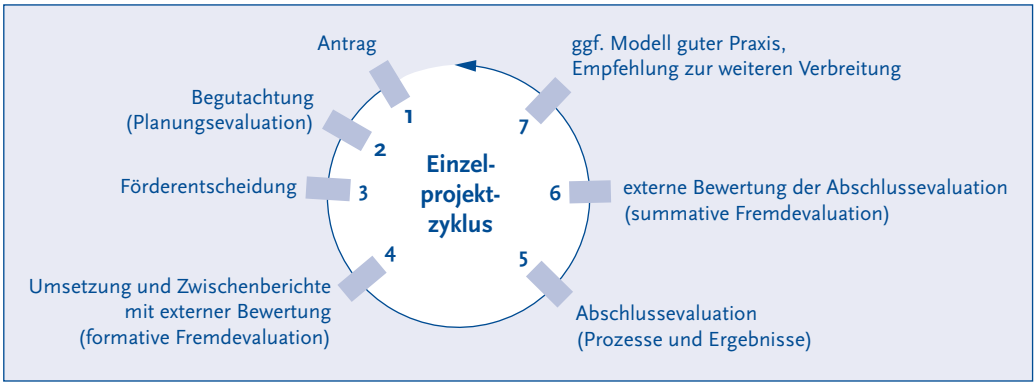


Tabelle 1
Evaluationsdimensionen und betrachtete Einzelaspekte bei der Evaluation der Gesamtinitiative Gesund.Leben.Bayern.

Evaluationsdimension	Einzelaspekte
Konzept	Schwerpunktsetzung der Initiative Leitprinzipien und Förderkriterien Qualitätsmanagementkonzept
Strukturen	Personal, Infrastruktur und Aufgabenverteilung bei der Abwicklung der Initiative Finanzielle Ausstattung der Initiative und Verteilung auf die einzelnen Schwerpunktbereiche Instrumente zum Qualitätsmanagement und zur Evaluation
Prozesse	Antrags- und Förderentscheidungsprozesse Umsetzung der Förderkriterien Umsetzung des Qualitätsmanagements
Ergebnisse	Health Impact – Gesundheitsziele Society Impact – Öffentlichkeitsarbeit und Agendasetting Science Impact – Beitrag zur Evidenzbasis in der GFP Practice and Policy Impact – Beitrag zur Qualitätskultur, Modellidentifizierung und ggf. Ausweitung, rechtliche Maßnahmen

Evaluationsmethoden

Die Evaluation der Gesamtinitiative war als externe, begleitende Evaluation angelegt und wurde von einem wissenschaftlichen Institut über die Jahre 2005 bis 2008 durchgeführt. Im Rahmen der externen Evaluation wurden begleitend zum Fortschritt der Initiative für die Jahre 2006, 2007 und 2008 Zwischenberichte zum Stand der Projektförderung erstellt, die auf den Projektdokumentationen, Zwischen- und abschließenden Projektberichten beruhten und vor allem einer Bewertung der Einzelprojektförderung dienten. Die abschließende Evaluation der Gesamtinitiative umfasste folgende Erhebungs- und Analyseschritte:

- Inhaltliche Bewertung des Konzepts der Initiative im Sinne einer fachlichen Stellungnahme. Basis der Bewertung war eine kritische Reflektion des Konzepts unter Berücksichtigung des aktuellen wissenschaftlichen Standes im Feld der GFP sowie anerkannter nationaler und internationaler Praxis.
- Aggregierte Analyse und Bewertung der Daten und Ergebnisse aus den Einzelprojektdokumentationen und Evaluationsberichten der Projektnehmer.
- Standardisierte Befragung der Projektnehmer sowie der abgelehnten Antragsteller.
- Semistandardisierte Datenerhebung bei den an der Abwicklung der Initiative beteiligten Verantwortlichen und Mitarbeitern (Befragung, Dokumenteneinsicht und -analyse).

Für die Ergebnisevaluation der Initiative unter den Aspekten Health Impact, Society Impact, Science Impact und Practice and Policy Impact wurden die folgenden Analysen angewandt:

Health Impact – Gesundheitsziele:

- Betrachtung der Verteilung der Projekte auf die Schwerpunkthandlungsfelder und Zielgruppen der Initiative
- Betrachtung der Reichweite (erreichte Personenzahl) und Impact (Wirkungen/Effekte) der einzelnen Projekte.

Eine Aggregation der gesundheitsbezogenen Einzelergebnisse der geförderten Projekte war nicht möglich, da die einzelnen Ansätze, Zieldimensionen und Indikatoren der Projekte zu unterschiedlich waren. Eine bevölkerungsbezogene Erhebung bzw. Monitoring zu Zielindikatoren der Initiative (z. B. Raucherquoten unter Jugendlichen, Adipositasprävalenz im Kindesalter) war im Rahmen der Initiative nicht vorgesehen.

Society Impact – Öffentlichkeitsarbeit und Agendasetting:

- Auswertung der Pressearbeit zur Initiative an sich sowie zu geförderten Projekten und Auswertung der Nutzung der Internetseite der Initiative
- Analyse des teilnehmenden Akteursspektrums im Bereich der Projektförderung und deren Netzwerkbildung
- Auswertung der geografischen und Setting-bezogenen Reichweite der geförderten Projekte.

Science Impact – Beitrag zur Evidenzbasis in der GFP:

- Bewertung von Qualität und Aussagen der Evaluationen der geförderten Einzelprojekte
- Auswertung der Anzahl der wissenschaftlichen Veröffentlichungen.

Practice and Policy Impact – Beitrag zur Qualitätskultur, Modellidentifizierung, Umsetzung von Vereinbarungen bzw. Rechtsetzung:

- Auswertung der Befragung der Projektnehmer zu Aspekten des Qualitätsmanagements
- Aggregierte Betrachtung der dokumentierten Qualitätssicherung bzw. -entwicklung und der Evaluationsgüte geförderter Projekte
- Identifizierung von Modellen guter Praxis anhand standardisierter Kriterien und Identifizierung erfolgreich ausgeweiteter Projekte
- Betrachtung umgesetzter Gesetzesinitiativen bzw. freiwilliger Vereinbarungen mit Bezug zur Initiative.

Verknüpfung der Evaluationsebenen

Die Evaluationen der Einzelprojekte ließen zum einen Aussagen zu den einzelnen Projekten zu. Zum anderen erlaubte die aggregierte Betrachtung der Einzelprojektevaluationen Aussagen zur Gesamtinitiative. Evaluationsfragen, über die die Evaluationsebenen »Einzelprojekte« und »Gesamtinitiative« verknüpften waren, waren z. B.:

- Konnten die Schwerpunkte der Initiative in der Einzelprojektförderung abgebildet werden? (Prozessevaluation und Ergebnisevaluation des Health Impacts der Gesamtinitiative)
- Konnten die Förderkriterien wie z. B. Evidenzbasierung, Evaluation, Settingansatz etc. angemessen umgesetzt werden? (Prozessevaluation der Gesamtinitiative)
- War das begleitende Qualitätsmanagement, d. h. die qualitätsorientierte Projektförderung und die damit verbundenen einzelnen Bewertungsschritte im Einzelprojektzyklus, leistbar und wurde es von den Projektnehmern akzeptiert? (Prozessevaluation und Ergebnisevaluation des Practice and Policy Impacts der Gesamtinitiative)
- Konnten die geförderten Projekte einen Beitrag leisten zu den Gesundheits- und nebengeordneten Zielen der Initiative? (Ergebnisevaluation des Health, Practice and Policy Impacts der Gesamtinitiative)

3 Hauptergebnisse der Evaluation von Gesund. Leben.Bayern.

In den Jahren 2006, 2007 und 2008 wurde ein Zwischenbericht zum Stand der Projektförderung sowie im Jahr 2010 der Ergebnisbericht zur Gesamtevaluation der Initiative vorgelegt.

Konzept, Leitprinzipien, Förderkriterien und Qualitätsmanagementansatz der Initiative wurden insgesamt positiv gewertet. Kritisch gesehen wurde die Diskrepanz zwischen der strategischen Zielsetzung der Initiative einerseits und der reaktiven Ausrichtung der Projektförderung in Abhängigkeit der Antragslage andererseits, was dazu führte, dass die Schwerpunktaktionsfelder der Initiative nicht durchgängig schwerpunktmäßig in der Projektförderung abgebildet wurden.

Science und Practice Impact: Alle im Zeitraum der Evaluation geförderten Projekte konnten plangemäß bzw. mit angebrachten Modifikationen abgewickelt werden. Dies bestätigt den zugrunde liegenden Projektauswahlprozess. Trotz der großen Bemühungen um die Projektevaluation lagen nur bei einem kleinen Anteil der Projekte qualitativ gute Evaluationen mit belastbaren Aussagen vor. Die besten Nachweise von Effektivität konnten im Handlungsfeld Ernährung und Bewegung erbracht werden. Insgesamt waren damit der Beitrag zur

Evidenzbasis wie auch die Möglichkeit zur Identifizierung und Empfehlung von Projekten als Modelle guter Praxis eingeschränkt, da der Wirkungsnachweis ein wichtiges Auswahlkriterien für Gute Praxis-Modelle darstellte. Entsprechend konnten im Evaluationszeitraum nur einzelne Projekte als modellhaft identifiziert werden. Obwohl knapp die Hälfte der Projektnehmer angab, durch die Antragstellung bei Gesund.Leben.Bayern. zur Auseinandersetzung mit Qualitätsaspekten angeregt worden zu sein, ließ die Bereitschaft zur Qualitätsentwicklung nach Antragstellung nach. Die Hilfestellungen zur Qualitätsentwicklung im Rahmen der Initiative wurden begrüßt (Workshop, Manuale, individuelle Beratung in kleinem Rahmen), weitere Hilfestellungen, vor allem eine Projektbegleitung im Sinne eines Mentoring, wären erwünscht gewesen.

Health and Policy Impact: Zu den Auswirkungen auf den Gesundheitszustand der Bevölkerung auf Bayern weiter Ebene konnten aus der Evaluation aus vielerlei Gründen, die im Folgenden diskutiert werden, keine unmittelbaren Aussagen abgeleitet werden. Gesetzliche Regelungen wurden erfolgreich im Bereich Rauchen auf den Weg gebracht, mit anfänglich einem Rauchverbot in Kindertagesstätten (BayKiBiG, Dezember 2005), dann in Schulen (BayEUG, August 2006) und zuletzt dem umfassenden Nichtraucherschutzgesetz (GSG, August 2010).

4 Diskussion der Evaluationsergebnisse

Die Evaluation von Gesund.Leben.Bayern. wurde bereits bei der Konzeption der Initiative berücksichtigt, in ihren Feinheiten jedoch erst in deren Verlauf entwickelt. Die Verbindung des von Anfang an bestehenden Evaluationsauftrags mit der schrittweise sich vollziehenden Entwicklung des Evaluationskonzepts bot die notwenige Flexibilität auf sich entwickelnde, nicht unbedingt vorhersehbare Eigenschaften und Dynamiken der Initiative einzugehen und ein kontinuierliches Lernen des Initiativenteams mit einzubeziehen und kam der Evaluation insgesamt zu Gute. Insofern waren die Evaluation wie auch die Initiative lernende Systeme.

Die Verknüpfung der beiden Evaluationsebenen »Einzelprojekte« und »Gesamtinitiative« im Rahmen der Gesamtevaluation kann als gelungen gelten. Die aggregierte Betrachtung der Einzelprojektförderung konnte vor allem zu Aspekten der Struktur- und Prozessevaluation beitragen. Insgesamt betrachtet liegen die Stärken der Evaluation der Gesamtinitiative vor allem in den Bereichen Konzept-, Struktur- und Prozessqualität. Die hier gewonnenen Einsichten geben wertvolle Hinweise für die weitere Ausgestaltung bzw. Abwicklung dieses und ähnlicher Vorhaben. Mit Blick auf die Ergebnisevaluation konnten auch zum Society, Science, Practice and Policy Impact hilfreiche Erkenntnisse gewonnen werden. Eine vertieftere Betrachtung dieser Bereiche wäre in Einzelaspekten wünschenswert gewesen. Hierzu ist eine weiterführende Indikatoren- und Methodenentwicklung vonnöten, die im Rahmen der Evaluation der Initiative nicht geleistet werden konnte.

Was die gesundheitsbezogenen Ergebnisse der Initiative anbelangt, können diese prinzipiell im Rahmen der Begrifflichkeiten "Health Impact", »Zielerreichung« bzw. »Wirksamkeit« der Initiative betrachtet und diskutiert werden. Jede dieser drei Betrachtungsweisen setzt einen anderen Schwerpunkt, hat ihre eigenen Voraussetzungen und Anforderungen an Untersuchungsmethoden und -design und erlaubt als Ergebnis Aussagen auf unterschiedlichen Ebenen: beim Health Impact liegt der Fokus der Betrachtung auf Veränderungen gesundheitlicher Parameter, bei der Zielerreichung findet die Betrachtung von Veränderungen in Beziehung zu den gesetzten Zielen statt, während die Debatte um die Wirksamkeit auf die Initiative als Intervention mit kausal zuzuordnenden Effekten fokussiert.

Eine *kausale* Betrachtung von Änderungen bevölkerungsbezogener Gesundheitsparameter im Bezug zur Gesundheitsinitiative war nicht Gegenstand der Gesamtevaluation und wurde entsprechend im Evaluationsdesign nicht berücksichtigt. Aussagen zum *Health Impact*, d. h. zu Veränderungen bevölkerungsbezogener Gesundheitsparameter, sind aus der Evaluation der Initiative aus verschiedenen Gründen nicht direkt ableitbar. Von der Ebene der Einzelprojekte aus betrachtet, war festzustellen, dass erst wenige der Einzelprojekte eine belastbare Ergebnisevaluation aufwiesen. Die vorhandenen Ergebnisse waren aufgrund der Heterogenität der Untersuchungsansätze nicht auf einer übergeordneten Ebene aggregierbar. Auf Bevölkerungsebene fand im Rahmen der Gesamtevaluation keine eigene Datenerhebung zu gesundheits-

bezogenen Parametern statt. Angesichts der zeitlich und oft auch räumlich begrenzten Natur der durch die Initiative geförderten Projekte sowie des relativ kurzen Evaluationszeitraums von 2005 bis 2008 sind Bayern weite Veränderungen gesundheitsbezogener Indikatoren nicht zu erwarten. Anders ist das Potenzial für mögliche bevölkerungsbezogene Effekte bei den im Verlauf der Initiative umgesetzten rechtlichen Maßnahmen zum Nichtraucherschutz. Hierzu fanden bisher keine dedizierten Datenerhebungen statt. Die Betrachtung der *Erreichung gesundheitlicher Ziele* leidet zum derzeitigen Zeitpunkt analog zur Betrachtung des Health Impacts an der fehlenden Datengrundlage. Das Fehlen von explizit formulierten Gesundheitszielen für einige der Schwerpunkthandlungsfelder der Initiative ist ebenfalls nachteilig.

5 Weiterführende Betrachtungen zur Evaluation komplexer Interventionen

Viele Maßnahmen der GFP können hinsichtlich ihres Ansatzes bzw. der Wirkungspfade als komplex bezeichnet werden. Das UK Medical Research Council definiert komplexe Interventionen als Interventionen mit mehreren interagierenden Komponenten, wobei sich Komplexität jedoch auch hinsichtlich anderer Dimensionen wie z. B. der Outcomes manifestieren kann (Craig et al. 2008). Varianten dieser Definition sind Beschreibungen komplexer Interventionen als »Interventionen mit verschiedenen unabhängigen oder aufeinander bezogenen Maßnahmen« (Shepperd et al. 2009), »Interventionen mit variablen oder schlecht zu definierenden Komponenten« (Wolf et al. 2012) oder »Interventionen, die größeren Schwankungen unterworfen sind als ein Medikament« (Campbell et al. 2000). Während dies alles auch für komplexe Interventionsprogramme wie die Gesundheitsinitiative Gesund.Leben.Bayern. zutrifft, wird im Kontext der vorliegenden Betrachtungen der Bezugsrahmen enger gewählt und komplexe Interventionen als Interventionsprogramme wie z. B. die bayerische Gesundheitsinitiative verstanden, die mehrere Projekte und Einzelaktivitäten unter einer übergeordneten strategischen Zielsetzung bündeln und oft verschiedene Settings und Themen, unterschiedliche Akteure, Sektoren und konzeptionelle Ansätze vereinbaren und sich meist auf größere Populationen bzw. Bevölkerungsteile beziehen. Neben der Eigenschaft der Multikomponenten-Intervention sind also auch die Eigenschaften einer Mehrebenen-Intervention und das Vorhandensein multipler Zielsetzungen bzw. Outcomes bezeichnend. Aus der Evaluation der Initiative Gesund. Leben.Bayern. lässt sich verallgemeinernd schlussfolgern, dass, so wie sich derartig komplexe Interventionen von – ggf. auch komplexen – Einzelmaßnahmen unterscheiden, auch die Evaluation dieser komplexen Interventionen in mehrfacher Hinsicht von der Evaluation einzelner Maßnahmen differiert. Es ergeben sich folgende Besonderheiten bei der Evaluation solch komplexer Interventionen:

Evaluationsgegenstand: Zu berücksichtigen sind mindestens zwei Ebenen, die des Gesamtprogramms mit übergeordneten Zielsetzungen als tragender und gestaltender Rahmen sowie die Ebene der Einzelmaßnahmen. Die beiden Ebenen sind miteinander verknüpft, woraus folgt, dass sie nicht in Isolation betrachtet werden können.

Evaluationszweck: Auch bei der Evaluation komplexer Interventionen können sowohl formative (gestaltende) als auch summative (bilanzierende) Zwecksetzungen verfolgt werden. Aufgrund der strategischen Ausrichtung, der oftmals politischen Anbindung, des meist beträchtlichen finanziellen Volumens und der öffentlichen Sichtbarkeit komplexer Interventionsprogramme besteht oft ein Primat der politischen, bilanzierenden Bewertung. Diese kann mit dem Ziel, Transparenz, Legitimation, öffentliche Rechenschaft und/oder eine Entscheidungsgrundlage zur Programmsteuerung herzustellen, verfolgt werden. Hierbei können konzeptionelle, strukturelle, Prozess- sowie Ergebnisaspekte von Wichtigkeit sein, deren jeweilige Gewichtung vom Einzelfall abhängt. Zwischen dem politischen Anspruch und der wissenschaftlich-methodischen Machbarkeit vor allem einer Ergebnisevaluation sowie der darauf aufbauenden politischen und wissenschaftlichen Bewertung kann ein Spannungsverhältnis bestehen. Eine grundsätzliche Klärung der Erwartungshaltungen im Vorfeld ist ratsam.

Evaluationsherangehen und Fragestellungen: Die Evaluation einer komplexen Intervention als Gesamtprogramm geht über eine Addition der Evaluationen der Einzelmaßnahmen hinaus. Zum einen sind zusätzliche, qualitativ andere Evaluationsaspekte und -fragestellungen einzube-

ziehen. Diese ergeben sich in Abhängigkeit vom Programm. Mögliche Aspekte sind z. B. die strategische Koordination der Einzelaktivitäten, Vernetzungsqualität, Herausbildung gemeinsamer Präventions- oder Qualitätskulturen oder die Evaluation von Synergien oder Antagonismen zwischen Teilen und Ebenen. Eine aggregierte Betrachtung der Einzelmaßnahmen unter diesen Gesichtspunkten trägt zur übergreifenden Bewertung des Gesamtprogramms bei und spiegelt die Verknüpfung der beiden Evaluationsebenen wider. Zum anderen kann der Ergebnisnachweis, vor allem was gesundheitsbezogene Ergebnisse betrifft, auf Ebene des Gesamtprogramms nicht einfach über die Wirkungsnachweise der Einzelmaßnahmen erbracht werden, d. h. die Wirkungen der einzelnen Maßnahmen können nicht einfach aufaddiert werden. Die Untersuchung der gesundheitsbezogenen Auswirkungen des Gesamtprogramms kann auf Ebene der Untersuchung des Health Impacts, der Zielerreichung oder der Wirkungsevaluation im Sinne der Effektivität des Programms stattfinden. Für Ersteres (Health Impact) kann ein Monitoring relevanter Zielgrößen dienen, für Zweiteres (Zielerreichung) die Betrachtung der Monitoringergebnisse in Bezug zu gesetzten Zielen. In Verbindung mit einer öffentlichen Diskussion dieser Ergebnisse kann einer politischen Evaluationszielsetzung mit Schaffung von Transparenz und öffentlicher Legitimation weitgehend Rechnung getragen werden. Für Letzteres (Effektivitätsnachweis) ist ein über ein Monitoring hinausgehender Evaluationsansatz zur kausalen Zuordnung von Veränderungen zum Programm erforderlich. Die Entwicklung tragfähiger Konzepte zur Wirkungsevaluation komplexer Interventionen ist Gegenstand nationaler sowie internationaler Debatten und Bemühungen (z. B. Campbell et al. 2000; Craig et al. 2008; Mühlhauser et al. 2011), wobei jedoch der Begriff »komplexe Intervention« in der Literatur zum Teil sehr unterschiedlich ausgelegt wird und für den Bereich komplexer Public Health-Programme bewährte epidemiologische und statistische Herangehensweisen bislang nicht zur Verfügung stehen.

Zusammenfassend ergibt sich für die Evaluation komplexer Interventionsprogramme in der GFP das Bild eines Methodenmixes aus quantitativen und qualitativen Ansätzen mit Integration der Ergebnisse der verschiedenen Evaluationsebenen. Die Evaluation der Gesundheitsinitiative Gesund.Leben. Bayern. mit Einsatz von (quasi)experimentellen Designs, Verlaufsbeobachtungen und qualitativen Methoden auf der Einzelprojektebene sowie quantitativen und qualitativen Auswertungen auf der Ebene des Gesamtprogramms veranschaulicht dies.

6 Resümee

Insgesamt betrachtet konnte die Gesundheitsinitiative Gesund.Leben.Bayern. als komplexe Intervention in vielerlei Hinsicht erfolgreich evaluiert werden. Die die ganze Initiative durchziehenden Leitprinzipien Qualität und Evaluation, die Einplanung einer Evaluation sowohl auf Einzelmaßnahmen- wie auch Gesamtinitiativenebene von Anfang an und die angemessene finanzielle Ausstattung der Evaluationsbemühungen waren wesentliche Voraussetzungen hierfür. Gelungen ist die Verknüpfung der beiden Evaluationsebenen Einzelmaßnahmen und Gesamtprogramm, aussagekräftige Ergebnisse liegen vor allem im Bereich der Konzept-, Struktur- und Prozessevaluation vor. Die Ergebnisevaluation bleibt insbesondere im Bereich der gesundheitsbezogenen Auswirkungen auf Bevölkerungsebene eine Antwort schuldig. Da die Evaluation komplexer Interventionsprogramme ebenso sehr ein politisches wie ein wissenschaftliches Unterfangen ist, ist zu diskutieren, ob im Hinblick auf die Ergebnisevaluation einem Monitoring entsprechender Zielparameter in Verbindung mit einer öffentlichen Diskussion der Monitoringergebnisse nicht ein ebenso hoher Stellenwert zukommt wie einer wissenschaftlichen Wirkungsevaluation. Beides wäre wünschenswert, doch das Fehlen bewährter methodischer Verfahren für Letzteres sollte nicht davon abhalten, zumindest Ersteres als gute Praxis für komplexe Interventionsprogramme zu etablieren.

Literatur

Campbell M, Fitzpatrick R, Haines A et al. (2000) A framework for development and evaluation of RCTs for complex interventions to improve health. BMJ 321 (7262): 694–696

Craig P, Dieppe P, Macintyre S et al. (2008) Developing and evaluating complex interventions: new guidance. UK Medical Research Council www.mrc.ac.uk/complexinterventionsguidance (Stand: 07.06.2012)

Loss J, Eichhorn C, Reisig V et al. (2007) Qualitätsmanagement in der Gesundheitsförderung. Entwicklung eines multidimensionalen Qualitätssicherungsinstruments für eine landesweite Gesundheitsinitiative. Prävention und Gesundheitsförderung 2 (4): 199–206

Mühlhauser I, Lenz M, Meyer G (2011) Entwicklung, Bewertung und Synthese von komplexen Interventionen – eine methodische Herausforderung. Z. Evid. Fortbild. Qual. Gesundheitswesen (ZEFQ) 105 (10): 751–761

Reisig V, Nennstiel-Ratzel U, Loss J et al. (2007) Das Evaluationskonzept der Bayerischen Gesundheitsinitiative »Gesund.Leben.Bayern.«. Prävention 30 (4): 116–119

Shepperd S, Lewin S, Straus S et al. (2009) Can We Systematically Review Studies That Evaluate Complex Interventions? PLoS Med. 6 (8): e10000086. doi:10.1371/journal.pmed.10000086

Sturm H, Weilmeier C (2004) Bürgergutachten für Gesundheit. Gesellschaft für Bürgergutachten München und Landshut (Hrsg) www.stmug.bayern.de/gesundheit/aufklaerung_vorbeugung/buergergutachten/doc/bgg01.pdf (Stand: 07.06.2012)

Wildner M, Kuhn J, Caselmann WH et al. (2007) Berichte aus den Ländern. Bayern. In: Gesellschaft für Versicherungswissenschaft und -gestaltung e. V. (Hrsg) Gesundheitsziele im Föderalismus – Programme der Länder und des Bundes. Nanos Verlag oHG, Bonn, S 21–33

Wildner M, Nennstiel-Ratzel U, Reisig V et al. (2006) Schwerpunkte der Prävention und Gesundheitsförderung in Bayern. Eine empirische Untersuchung. Prävention und Gesundheitsförderung 1 (3): 149–158

Wolf K, Schiffner R, Rheinberger P (2012) Bewertung komplexer Interventionen im Gemeinsamen Bundesausschuss. Meeting Abstract. Komplexe Interventionen – Entwicklung durch Austausch. 13. Jahrestagung des Deutschen Netzwerks Evidenzbasierte Medizin. Hamburg, 15.–17.03.2012. Düsseldorf: German Medical Science GMS Publishing House; 2012. Doc12ebm016 www.egms.de/static/en/meetings/ebm2012/12ebm016.shtml (Stand: 06.07.2012)

Evaluation der Netzwerke Gesunde Kinder im Land Brandenburg – Einige Erkenntnisse aus der praktischen Evaluation komplexer Interventionen

Wolf Kirschner, Nicole Rabe

1 Einführung

Die Evaluationsforschung hat in Deutschland im Vergleich zu den USA oder Großbritannien keine längere Tradition. Erst in den letzten zwei Jahrzehnten ist die Nachfrage nach Evaluationsmaßnahmen gestiegen. Es sollen möglichst nur solche medizinischen und sozialen Interventionen in breite Anwendung kommen, für die der Nachweis ihrer Wirksamkeit und Wirtschaftlichkeit erbracht worden ist. Zunächst ist einschränkend zu konstatieren, dass dieses Evidenzpostulat bis heute in der medizinisch/sozialen Versorgung nicht durchgehend umgesetzt ist und nicht Alles, was den Titel »Evaluation« trägt, beinhaltet tatsächlich auch Evaluationsforschung. Dennoch stellt sich die Frage nach den Standards für einen wissenschaftlichen Nachweis der Wirksamkeit und Wirtschaftlichkeit von Interventionsmaßnahmen. Es kann auch nicht verwundern, dass mit der Institutionalisierung des noch recht jungen Interventionsfeldes »Frühe Hilfen« in Deutschland die wissenschaftliche Debatte über angemessene Evaluationsdesigns (Schmacke 2009) wieder aufflammt.

Im Kern geht es um die Frage, ob zur Überprüfung der Evidenz der Einsatz der als Goldstandard deklarierten Randomised Control Trials (RCT) unabdingbar sind oder ob auch andere Evaluationsdesigns – auch im bewussten Verzicht auf Kontrollgruppen – belastbare Evaluationsergebnisse erbringen können. In einer entsprechenden Expertise des Nationalen Zentrums »Frühe Hilfen« vertreten Lengning und Zimmermann (2010; 2009) dezidiert und rigoros die erste Position, während z. B. Ziegler (2010), Boettcher et al. (2009) und Elkeles (2006) erhebliche Zweifel an der Angemessenheit von RCT äußern und für eine alternative Wirkungsforschung als »Prozess-Mechanismus-Forschung« bzw. für »realistische quasiexperimentelle Wirkungsanalysen« plädieren. Im vorliegenden Beitrag werden nach einer Erörterung der politischen und programmatischen Rahmenbedingungen im Interventionsfeld das Design der Evaluation der Tätigkeit der Netzwerke Gesunde Kinder (NGK) und auch erste Erkenntnisse und Ergebnisse beschrieben. Wir denken, dass sich aus dem praktischen, empirischen Zugang einige Hinweise zur Frage angemessener Evaluationsstandards in diesem Handlungsfeld ableiten lassen.

2 Die politische Implementation von Interventionen als Reaktion auf soziale und politische Problemkonstellationen

Die seit Beginn des Jahrhunderts zunehmenden oder zunehmend wahrgenommenen Fälle von Kindeswohlgefährdung sowie Straftaten wie Mord und Totschlag an Kindern werfen Fragen hinsichtlich Qualität, Effektivität und Effizienz der Arbeit der Jugendhilfe bzw. der Regulation des gesamten Handlungsfeldes auf. Die Politik kann auf ein gesellschaftliches Problem (Abweichung eines Ist- von einem Sollzustand) durch eine reformerische Neugestaltung der Zuständigkeiten und Aufgaben und/oder durch die Schaffung neuer und zusätzlicher Angebote und Instrumente reagieren. Sowohl die Bundesregierung als auch die Landesregierungen setzen seit 2005/2006 bei weitgehender Aufrechterhaltung bestehender Zuständigkeiten im Gesundheits- und Jugendbereich[1] auf soziale, präventive Interventionen im Problemfeld, die insgesamt mit dem Begriff der »Frühen Hilfen« umrissen werden, auch wenn diese z. T. recht unterschiedlich konzipiert sind. Während noch in den 80er- und 90er-Jahren des letzten Jahrhunderts soziale Interventionen in Deutschland nur selten und dann meist in Form von Selbstevaluationen untersucht wurden, ist die Evaluation entsprechender Maßnahmen – geartet wie auch immer – seit einigen Jahren zu einem festen Bestandteil administrativen Handelns und zunehmend auch zur Voraussetzung administrativer Mittelbewilligung für Interventionen geworden.

1 Erst im Januar 2012 wurden die Rechtsgrundlagen mit dem Kinderschutzgesetz neu kodifiziert.

3 Charakteristika der NGK im Kontext der Projekte »Frühe Hilfen«

Die Bundesregierung hat 2006 das Programm »Frühe Hilfen für Eltern und Kinder und soziale Frühwarnsysteme« aufgelegt, mit dem in Ländern und Kommunen risikoorientierte Modellprojekte initiiert, durchgeführt und evaluiert werden sollen. Die Landesregierung in Brandenburg hat unabhängig davon in einem Maßnahmenpaket für Familien- und Kinderfreundlichkeit vom Dezember 2005 insgesamt 61 Maßnahmen benannt. Zu diesen Maßnahmen gehören an vorderster Stelle die NGK mit einer zunächst geplanten Laufzeit von 2006 bis 2008. Gefördert wurden die NGK vom Ministerium für Arbeit, Soziales, Gesundheit und Familie. Die grundsätzlichen Charakteristika der NGK sind:

- Vernetzung aller Akteure und Leistungen von Gesundheit (SGB V) und Jugendhilfe (SGB VIII)
- Anbindung der Netzwerkorganisation und -koordination an Institutionen auch jenseits der Gesundheits- oder Jugendämter, d. h. z. B. an Kliniken oder andere Einrichtungen (Vereine, Wohlfahrtsverbände)
- Gehstruktur mit aufsuchenden personalen Interventionen vornehmlich durch ehrenamtliche Patinnen/Paten (an einem Standort im 1. Lebensjahr der Kinder durch Hebammen)
- Bevölkerungsweiter, nicht risikogruppenspezifischer Ansatz.

Die NGK in Brandenburg unterscheiden sich von anderen Projekten »Frühe Hilfen« somit v. a. in der nichtrisikogruppenspezifischen Ausrichtung, in einer unterschiedlich möglichen Rolle des Jugendamtes im Akteursnetzwerk und in der aufsuchenden ehrenamtlichen Arbeit. Alle drei Charakteristika resultieren letztlich aus dem Umstand, dass im Jahr 2005 das Klinikum Niederlausitz dem Ministerium ein ausgearbeitetes und bereits lokal umgesetztes Interventionskonzept (EKIB = Entwicklung von Kindern in Beziehung) vorstellte, das eben diese Kriterien aufwies (Reinisch 2011). Erwähnt werden muss darüber hinaus, dass Ministerpräsident Platzeck Mitte 2004 bei seinem Besuch in Finnland auf das Neuvola Projekt (Sass o. J.) zur Betreuung von Familien und Kindern aufmerksam wurde, welches in weiten Teilen zu dem Interventionskonzept aus Niederlausitz Parallelen aufwies. Das NGK war damit von Beginn an ein politisch sehr hoch angesiedeltes Interventionsprojekt. Insgesamt versteht sich das NGK nicht primär als Programm des Kinderschutzes, vielmehr als Programm zur Förderung der gesundheitlich/sozialen Entwicklung von Kindern und ihrer Familien.

4 Projektspezifische Ausgangsbedingungen bei der Konzeption des Evaluationsdesigns im Jahr 2007

Die ersten drei NGK nahmen im Juni 2006 (Lauchhammer), im Dezember 2006 (Eberswalde) und im März 2007 (Havelland) mit der abgeschlossenen Schulung der ersten Patengruppen ihre Arbeit auf. Die Evaluation wurde im März 2007 beschränkt ausgeschrieben und mit einer Laufzeit zunächst bis Ende 2009 im September an uns beauftragt. Der Start einer Intervention vor der Festlegung des Evaluationsdesigns und der Entwicklung der Evaluationsinstrumente ist in der Praxis der Evaluationsforschung nicht selten. Die nicht mögliche Teilnahme eines Evaluationsteams an der (abschließenden) Konzeption eines Interventionsprojektes hat verschiedene politisch-institutionelle und administrative Ursachen. Aus der Nichtteilnahme eines Evaluationsteams an der Konzeption eines Interventionsprojektes können regelmäßig drei Probleme resultieren:

1. Eine fehlende ex-ante Evaluation hinsichtlich der Ableitung eines Wirkungsmodells und daraus resultierenden konzeptionellen Defiziten in der Ableitung von empirisch messbaren Zielen.
2. Defizite und Mängel in den projektinternen Daten- und Informationssystemen, die für das interne Projektcontrolling und auch für die externe Evaluation verwendet werden.
3. Unklarheiten über die Aufgaben der Evaluation unter den Projektdurchführenden und ggf. mangelnde Akzeptanz der Evaluation.

Während die Akzeptanz der Evaluation unter den Projektdurchführenden sehr hoch war, traten die beiden erst genannten Probleme erwartungsgemäß auf. In den drei Projekten wurden ganz unterschiedliche und nichtabgestimmte Dokumentationssysteme über Akteure, Paten, Familien und Kinder geführt, die zunächst eine gemeinsame Auswertung nicht zuließen. Es dauerte ein Jahr bis eine inhaltlich und technisch abgestimmte Datenbank (AmbuCare) zur Verfügung stand und vergleichende Auswertungen erlaubte. Beim NGK handelt es sich um ein Programm, das interventionstheoretisch definiert werden kann als »der gezielte und effektive Einsatz von (wirksamen) Maßnahmen unter Verwendung definierter Ressourcen zur Erreichung bestimmter Ziele, die ein bestehendes und klar beschriebenes Problem vermindern oder beseitigen sollen«(Elkeles, Kirschner 2003). Die zentrale Frage der ex-ante Evaluation mit Blick auf die Ergebnisevaluation, d. h. die Programmwirkungen ist, ob mit den geplanten Interventionsinstrumenten (aufsuchende Beratung und Unterstützung) bestimmte intendierte Ziele in welchem Maße erreicht werden können. Dies setzt zunächst eine Definition der Ziele voraus, die in der Ausschreibung als »Verbesserung des somatischen und psychischen Gesundheitszustandes der Kinder« sowie der »Entwicklung günstiger Beziehungen in den Familien« noch eher allgemein definiert waren. Zur Frage der potenziellen Wirksamkeit der vorgesehenen Interventionen können prinzipiell die Ergebnisse der Sozialepidemiologie und der bisherigen Interventionsforschung auf dem Handlungsfeld »Frühe Hilfen« herangezogen werden. Sozialmedizinische Befunde zeigen, dass die Ursachen von Kindeswohlgefährdungen vielfältig sind, sich mangels hinreichend großer retrospektiver oder prospektiver Studien aber der Quantifizierung durch Assoziationsmaße, die für die Auswahl potenziell wirksamer Maßnahmen wünschenswert sind, weitgehend entziehen. Die Befunde aus der Evaluation von einigen Projekten im Kontext von Jugendhilfe und Gesundheit zeigen, dass entsprechende Interventionen Wirksamkeit haben können, auch wenn diese oft auf geringen Fallzahlen basieren und die Effekte eher gering sind (Lengning, Zimmermann 2009; Friese, Walter 2005). Insgesamt zeigt sich, dass es Hinweise auf die Wirksamkeit von Maßnahmen der aufsuchenden Familienarbeit gibt und positive Wirkungen prinzipiell plausibel erscheinen. Dies mag zunächst ernüchternd erscheinen. Allerdings lässt sich im weiten Feld sozialer Interventionen kaum ein Beispiel finden, welches kausalanalytisch unterlegt ist. Das Kriterium zur Überprüfung der möglichen Wirksamkeit von Maßnahmen in der ex-ante Evaluation ist somit in den meisten Fällen nicht die Korrelation oder gar die Kausalität. Letztlich basiert die Intervention also auf der Annahme, dass eine positive Wirkung auf die genannten Zielvariablen durch aufsuchende Familienarbeit plausibel erscheint, wobei die Aufgabe der notwendig weiteren Ausdifferenzierung der Zielvariablen an einen Arbeitskreis aus Vertretern des Interventions- und Evaluationsteams unter Federführung des für Familie zuständigen Ministeriums delegiert wurde.

5 Schwerpunktsetzung auf die Evaluation von Strukturen, Prozessen und Produkten unter Berücksichtigung von kurz- und mittelfristigen Wirkungsanalysen

Neben den genannten Ergebniszielen wurden in der Ausschreibung auch Struktur- und Prozessziele als Teilziele bzw. Voraussetzung zur Erreichung der Ergebnisziele wie folgt definiert:

1. Aufbau der Netzwerkorganisation, Vernetzung mit Akteuren, Gewinnung und Schulung von ehrenamtlichen Paten
2. Diffusion der drei Modellprojekte in möglichst alle Kreise und kreisfreien Städte im Land Brandenburg
3. Die Gewinnung von Familien und Kindern und die Akzeptanz der Netzwerkarbeit bei den Adressaten.

Das von uns vorgeschlagene Evaluationsdesign basierte auf der Evaluationsphilosophie unseres Instituts, wonach die Struktur- und Prozessevaluation im Vergleich zur Ergebnisevaluation mindestens gleichberechtigte, wenn nicht zunächst prioritäre Handlungsfelder der Evaluation sein müssen. Die vorschnelle Fokussierung auf Ergebniswirkungen ist abzulehnen. Rossi et al. (1992) u. a. formulieren drastisch: »Es ist offensichtlich eine Verschwendung von Zeit, Mühe und Ressourcen, die Wirkung eines Programms festzustellen, das keine

messbaren Ziele hat und nicht ordnungsgemäß implementiert wurde.«

5.1 Maßnahmen und Instrumente der Struktur- und Prozessevaluation

Im Rahmen der **Strukturevaluation** wurden unmittelbar nach Auftragserteilung folgende Aspekte bearbeitet:

- Beschreibung der sozioökonomischen Lage in allen Landkreisen und kreisfreien Städten
- Beschreibung der Bedarfslage aus der Sicht der Träger und Kommunen (hier v. a. Jugend- und Gesundheitsamt)
- Beschreibung der Projektziele an den Standorten
- Beschreibung der Netzwerkstrukturen (Netzwerkanalysen)
- Bewertung von Qualität und Eignung der einzelnen Angebote und Maßnahmen
- Analyse der Projektdokumentationssysteme.

Grundlage der Arbeiten waren neben Desk-Research v. a. qualitative Interviews mit den Netzwerkleitungen und Akteuren. Die Analysen zeigten mit Ausnahme der unzureichenden Dokumentationssysteme zusammenfassend eine gute Strukturqualität, einen als hoch eingeschätzten Bedarf für die Intervention, eine hohe Motivation der Akteure bei allerdings erheblichen sozioökonomischen Unterschieden, nicht nur zwischen den Interventionsregionen, sondern auch innerhalb einzelner Kreise. Es ist auch darauf hinzuweisen, dass die differenzierte Analyse der Sozialstruktur mit Hilfe des Sozialindexes von Einschülern nur auf der Grundlage der ausgebauten Gesundheitsberichterstattung im Land Brandenburg überhaupt möglich war (MASF o. J.). Ferner war zu konstatieren, dass die erforderliche und gewollte, aber vielfach ungewohnte Kooperation ganz unterschiedlicher Institutionen und Professionen teilweise auch konfliktreich war und sich nicht selten in Kommunikations- und Verständigungsproblemen ausdrückte.

Im Rahmen der **Prozessevaluation** wurde untersucht:

- Statistik der Teilnehmer und Abbrecher insgesamt und nach sozialer Lage (Programmreichweite)
- Statistik der Paten und Akteure
- Statistik der weiteren Maßnahmen der Netzwerke (z. B. Veranstaltungen, Schulungen)
- Analyse der medialen Berichterstattung und der Internetauftritte
- Analyse der getroffenen Maßnahmen zur Qualitätssicherung und der Netzwerkentwicklung
- Analyse der Diffusion der NGK.

Die Maßnahmen der Struktur- und Prozessevaluation wurden schrittweise auf die in Folge gegründeten nunmehr 17 NGK ausgeweitet. Hierbei trat das Problem auf, dass neue Netzwerke teilweise eine geringere Akzeptanz für die Evaluation und den Einsatz der Evaluationsinstrumente zeigten, was primär darauf zurückzuführen war, dass sie an deren Konzeption nicht beteiligt waren. In Abstimmung mit den Netzwerkleitungen, Koordinatorinnen, Patinnen/Paten und Vertretern des Ministeriums wurden nach den qualitativen Vorarbeiten bis Ende 2008 folgende quantitativen Evaluationsinstrumente entwickelt und in Einsatz gebracht, die überwiegend der fortlaufenden Struktur- und Prozessevaluation dienten, partiell jedoch auch für die Ergebnisevaluation Verwendung finden:

- Mütterfragebogen (ab 9. Monat des Kindes als Kohortenuntersuchung konzipiert)
- Patenfragebogen (einmalig nach abgeschlossener Schulung)
- AmbuCare Datenbank
- Prozessdokumentationsbogen (Quartalserhebung).

Der Entwicklungsprozess dieser Instrumente ist durch Probleme gekennzeichnet, wie sie jeder Evaluationsforscher aus der Praxis kennt. Die wissenschaftlich-instrumentellen Erkenntnisinteressen der Evaluatoren müssen in einen oft schwierigen Kompromiss gebracht werden, mit dem, was das Interventionsfeld hier zuzugestehen bereit ist. Hier sind nicht nur die bestehenden personellen und zeitlichen Ressourcenprobleme auf Seiten der

Interventionsteams zu nennen, vielmehr kommen gerade in Bezug auf die Fragebogeninstrumente oft alle Vorbehalte gegenüber der Notwendigkeit und Sinnhaftigkeit der empirischen Forschung sowie oft unbegründete datenschutzrechtliche Bedenken auf den Tisch. Im "collective bargaining" um das adäquate Design ist meist die möglichst geringe zeitliche Belastung der Teams, wie auch der Teilnehmer letztlich das Hauptentscheidungskriterium über den Umfang und damit auch den Inhalt der Instrumente.

5.2 Maßnahmen und Instrumente der Ergebnisevaluation

Eine Ergebnisevaluation von Interventionen kann bei »tragfähigen« Strukturen und Prozessen einerseits erst dann erfolgen, wenn Wirkungen auf den Gesundheitszustand und die Familienbeziehungen in relevantem Maße überhaupt sichtbar werden können, andererseits muss das vorgegebene Interventionsfenster bis zum 3. Lebensjahr des Kinder berücksichtigt werden. Die kurz- und mittelfristigen Wirkungen bis zum 3. Lebensjahr sollten im Mütterfragebogen[2] erhoben werden, der ab dem 9. Monat des Kindes erstmals und dann in jährlichen Abständen eingesetzt wird. In Bezug auf die gesundheitlichen Wirkungen wurden in Anlehnung an den KiGGS[3] Fragen zu folgenden Bereichen operationalisiert:

- Frühgeburt, Geburtsgewicht, Häufigkeit und Art gesundheitlicher Probleme in den ersten Lebenswochen
- Stillhäufigkeit und -dauer
- Subjektive Einschätzung des Gesundheitszustandes des Kindes
- Ausgewählte Krankheiten und Beschwerden jemals und in den letzten 12 Monaten
- Sprach-, Hör- oder Sehprobleme
- Körpergröße und -gewicht und Einschätzung der Kindesentwicklung
- Inanspruchnahme medizinischer/sozialer Leistungen inkl. spezieller Therapien in den letzten 3 und 12 Monaten.

In Bezug auf die Familienbeziehungen wurden folgende Fragen operationalisiert:

- Frage zur Zufriedenheit mit verschiedenen Lebensbereichen, darunter die Zufriedenheit mit der familiären Situation
- Fragen zu Dimensionen des elterlichen Erziehungsverhaltens nach Tschöpe-Scheffler und Niermann (2002) in einer reduzierten Form mit neun Items.

Beim Mütterfragebogen kam eine Fall-Kontroll-Studie z. B. in Form einer repräsentativen Bevölkerungsbefragung von Müttern mit gleichaltrigen Kindern in den Interventionsregionen v. a. aus Kostengründen nicht in Betracht.[4] Ein Vergleich der Teilnehmer war aber – zumindest über Randdaten – mit dem (regionalisierten) KiGGS-Datensatz möglich. Im Mittelpunkt der kurz- und mittelfristigen Wirkungsanalysen stand aber nicht primär die Frage nach den Unterschieden zwischen Teilnehmern und Kontrollen, sondern vielmehr zunächst die Frage, ob sich zwischen Kindern mit Müttern unterschiedlicher Schulbildung Unterschiede in den Zielvariablen ermitteln lassen. Die dahinter stehende Hypothese war, dass die aufsuchende Patenarbeit auf die sozialepidemiologisch bekannten schichtenspezifischen Morbiditäts- und Inanspruchnahmeprofile nivellierend wirkt.

Lag für die Analyse der mittelfristigen Wirkungen damit ein durchgängiger Vergleichsdatensatz nicht vor, so können die längerfristigen Wirkungen auf der Basis der einrichtungsbezogenen Daten der Kita-Untersuchungen und v. a. der Schuleingangsuntersuchungen auf hohen Fallzahlen belastbar untersucht werden. Hierzu wird in den Gesundheitsämtern die Teilnahme an den NGK codiert.[5] Die ersten Analysen hierzu können ab dem Jahr 2013 erfolgen. Diese Vergleichsmöglichkeiten erschienen uns in Bezug auf die Evaluationsqualität unter folgenden Gesichtspunkten überzeugend:

2 Der Mütterfragebogen enthielt auch Fragen zum Netzwerkzugang, zur Netzwerkzufriedenheit und zur Weiterempfehlungsbereitschaft.

3 Kinder- und Jugendgesundheitssurveys (KiGGS) des RKI

4 Darüber hinaus spielten auch kritische Überlegungen hinsichtlich der Repräsentativität, der Ausschöpfungsquote und der erwartbaren Selektivität dieser möglichen Vergleichsstichprobe eine Rolle.

5 Die Daten der Kita-Untersuchungen sind dafür allerdings nur eingeschränkt nutzbar, da diese den Sozialstatus der Eltern nicht beinhalten.

- wenn sich positive gesundheitlich-soziale Wirkungen ergeben, so sollten diese auch stabil und längerfristig d. h. auch im fünften oder sechsten Lebensjahr messbar sein
- selbst wenn die ärztlichen Befunde Qualitäts- und Validitätsprobleme aufweisen würden, ist nicht davon auszugehen, dass diese einen Bias im Vergleich von Teilnehmern und Nichtteilnehmern darstellen.

6 Bisherige Ergebnisse und Erkenntnisse

Mit Blick auf die Interventionsziele kann fast sechs Jahre nach dem Start des ersten Netzwerkes festgestellt werden, dass:

- die Diffusion der Netzwerke in alle Kreise und kreisfreien Städte bis auf drei Ausnahmen gelungen ist
- die Etablierung der Netzwerke mit tragfähigen Strukturen und Prozessen gut gelungen ist
- die Netzwerke in der Bevölkerung und unter Teilnehmern eine sehr hohe Wertschätzung aufweisen
- allerdings in den Interventionsregionen sehr deutliche Unterschiede in der Netzwerkteilnahme und Programmreichweite bestehen, die zwischen 40 % und unter 10 % variieren.[6]

Bereits der letzte Befund unterstreicht die Notwendigkeit einer konsequenten und realistischen Prozessevaluation. Bevor finale Ergebniswirkungen untersucht werden können, müssen hier zunächst die multiplen Ursachen der sehr unterschiedlichen Reichweiten ermittelt werden, denn die Annahme gleicher Wirksamkeit in den Regionen erscheint bei diesen Teilnahmedifferenzen nicht plausibel. Dies umso mehr, als die Analyse der Fragebogen (derzeit n=1.824) nach Schulabschluss der Mütter ein eindrucksvolles Bild über die weitere Heterogenität der Interventionspopulation mit einigen »Überraschungen« liefert. Wir fragen also zunächst nicht: "Does this program work?" 'but ask instead' "What works for whom in what circumstances and in what respects, and how?" (Pawson, Tilley o. J.).

6 Eine Zusammenfassung der Evaluationsergebnisse für das Jahr 2011 findet sich im Internet (MASF 2011)

Bei der sehr hohen Heterogenität der Population in der Soziodemografie und der Erwerbstätigkeit überraschen die geringen und nicht signifikanten Unterschiede bei den erfragten gesundheitlichen Indikatoren. Während das geringe Geburtsgewicht erwartungsgemäß unter Hauptschülerinnen signifikant höher ist, ist diese typische schichtspezifische Verteilung – im Gegensatz zu den KiGGS-Daten – beim aktuellen Gesundheitszustand der Kinder nicht zu beobachten. Auch wenn mögliche Artefakte nicht gänzlich auszuschließen sind, bestätigt dies die o. g. Hypothese. Andererseits werden auch bekannte Unterschiede bestätigt (Stillen, Kinderarztbesuche). Die unerwartet deutlich höhere Inanspruchnahme stationärer Leistungen unter Kindern mit Hauptschulmüttern, war Gegenstand umfangreicher Nachrecherchen.

Letztlich gibt es tragfähige Hinweise, dass diese nicht in erhöhter Morbidität begründet ist, sondern dass Netzwerkmütter dieser Gruppe vielmehr die »Krankenhausnähe der Netzwerke« verstärkt für die »schnelle« Inanspruchnahme nutzen. Auch zeigt sich, dass die Hauptschulmütter die Hilfen des Netzwerks beim Aufwachsen der Kinder deutlich positiver einschätzen.

7 Schlussfolgerungen

Einen aktuellen Beitrag zur kontroversen Diskussion über die Verwendung von RCTs bei sozialen Interventionen liefert Macintyre (2011). Darüber hinaus sind wir jedoch der Auffassung, dass die propagierte Anwendung von RCTs grundsätzlich auf Annahmen basiert, die auf diesem Handlungsfeld nicht gegeben sind. RCT sind standardisierte statistische Verfahren, die die Wirksamkeit einer klar definierten, »innovativen« medizinischen Intervention an eindeutig definierten und recht homogenen Populationen im Fall-Kontrollvergleich zur Standardtherapie oder Placebo mit definierten Stichprobenumfängen überprüfen. RCTs setzen voraus, dass aus Voruntersuchungen Ergebnisse zu den Effektgrößen der Wirksamkeit bekannt sind, da ohne diese eine exakte Planung der erforderlichen Stichprobengrößen in Bezug auf den α- und β-Fehler nicht erfolgen kann. Medizinische Innovationen bei Arzneimitteln und Medizinprodukten zielen regelmäßig auf die Optimierung von Gewinnen. RCTs stellen finanzielle Investitio-

Tabelle 1
Ausgewählte Variablen nach Schulabschluss der Mutter

	Gymnasium	Hauptschule	T-Wert
	n=585	n=469	
Alter der Mutter (Mittelwert/STABW) [n=571/448]	31,5/4,5	27,6/6,1	11,25
Derzeit (mehr als eine Stunde täglich) erwerbstätig	60,7%	30,5%	10,21
Geburtsgewicht <2.500 Gramm	5,8%	9,2%	2,04
Auf Geburt gut vorbereitet (Sehr, eher)	95,5%	90,2%	3,25
Kind gestillt (ja)	89,6%	74,5%	6,34
Bewertung Gesundheitszustand des Kindes (Sehr gut/gut)	9,5%	9,5%	0,43
Kind Probleme mit Sprechen, Hören, Sehen	7,0%	6,0%	0,17
Kind noch andere Krankheiten, Behinderungen, Unfälle	24,0%	19,8%	1,61
Kinderarztbesuche in den letzten 3 Monaten mind. 1-mal	91,5%	80,0%	5,27
Kind in letzten 12 Monaten stationär behandelt	22,3%	32,4%	3,62
Zufrieden mit familiärer Situation (Sehr/eher)	92,8%	87,4%	2,85
Netzwerk hat beim Aufwachsen des Kindes geholfen (Sehr stark/stark)	29,9%	51,4%	7,05
Alter des Kindes in Monaten (Mittelwert/STABW) [n=585/466]	18,7/8,8	17,8/8,5	1,54

nen mit Blick auf das Marketing und die Sicherheit eines neuen Produktes dar. Sie legitimieren sich als Forschungsaufwendungen in den ökonomischen Planungen der jeweiligen Unternehmen. RCTs sind bekannte, in der Beantragung, Begutachtung, Durchführung und Publikation durch "good clinical practice" in hohem Maße standardisierte und qualitätsgesicherte, allerdings sehr kostenintensive Verfahren.

a) Soziale Interventionen werden zur möglichen Verringerung eines gesellschaftlichen Problems politisch gewollt und implementiert und ihre Einführung, ihre Modifikation wie auch ggf. ihr Auslaufen sind politische Entscheidungen, bei denen Kriterien der Wirksamkeit oder gar Wirtschaftlichkeit der Interventionen nicht selten nachrangig sind, jedenfalls nicht zwingend gegeben sein müssen. Als Beispiel kann die Implementation der HIV-Präventionsstrategie Mitte der 1980er-Jahre genannt werden, die i. w. auf Plausibilitätsüberlegungen basierte.
b) Das Primat der Politik in der Initiierung sozialer Interventionen impliziert, dass Art und Inhalt der Interventionen immer auch von der Politik festgelegt werden und nicht prioritär wissenschaftlichen Gesichtspunkten und Begründugen folgen.[7] Evaluationsforschung hat in diesem Kontext die Aufgabe, das politische Handeln auf diesem Feld zu legitimieren, wobei die möglichen Inhalte und Designs der Evaluationen sehr unterschiedlich sein können.[8] Keinesfalls kann davon ausgegangen werden, dass die Politik »Gold-Standard Evaluationen« fordert oder gar bevorzugt. Viel stärker werden Inhalte und Designs der Evaluation durch die gegebenen finanziellen Ressourcen determiniert und oft auch limitiert.
c) Während die Akteure »klinischer Prüfungen« (Ärzte, Biometriker) mit der Intervention und der wissenschaftlichen Begleitung vertraut sind, gilt dies für die Akteure im Bereich sozialer Interventionen i. d. R. nicht. Zudem gibt es in diesem Handlungsfeld kein Standarddesign der Evaluationsforschung. Da die Evaluation letztlich von der Mit- und Zuarbeit der Interventoren abhängig ist, müssen diese für ein konkretes Evaluationsdesign gewonnen werden. Letzteres richtet sich dabei häufig nicht primär an Kriterien exzellenter wissenschaftlicher Qualität, sondern vielmehr an Kriterien der Praktikabilität.
d) Soziale Interventionen sind regelmäßig keine standardisierbaren Dienstleistungen oder Pro-

7 So unterscheiden sich die verschiedenen Projekte »Frühe Hilfen« gerade darin, ob sie sich als Maßnahmen des Kinderschutzes oder als Maßnahmen des vorsorgenden Sozialstaates (Schroder o. J.) verstehen.
8 Dies kann auch rein legitimatorische Evaluationen beinhalten.

dukte, die sich an homogene Gruppen richten, sie haben eher den Charakter von komplexen Modellversuchen, die oft multivariate und nicht selten auch diffuse Zielstellungen aufweisen. Über die Effektstärken ist oft wenig bis nichts bekannt, womit auch eine statistisch begründete Festlegung von Fallzahlen nicht möglich ist. Entsprechend erfolgt die »Fallzahlplanung« oft nur nach den vorhandenen personellen oder finanziellen Ressourcen.

Unsere erfahrungsbedingte Skepsis in Bezug auf die Adäquanz von RCTs bei sozialen und komplexen Interventionen bedeutet natürlich nicht, dass die Evaluation hier keine hohen Standards aufweisen müsste. Das Gegenteil ist der Fall. Bei allen methodischen Kontroversen gilt es immer, falsch-negative oder falsch-positive Evaluationsergebnisse mit hoher Sicherheit zu vermeiden. Vor diesen fatalen Ergebnissen sind aber auch RCTs nicht grundsätzlich gefeit. Evaluation "is a form of applied research, *not* performed for the benefit of science as such, but pursued in order to inform the thinking of policy makers, practitioners, program participants and public" (Pawson, Tilley 2009).

Die Evaluation der Netzwerke wird aus Mitteln des Ministeriums für Arbeit, Soziales, Frauen und Familie des Landes Brandenburg gefördert.

Literatur

Boettcher W et al. (2009) Soziale Frühwarnsysteme und Frühe Hilfen, Modelle, theoretische Grundlagen und Möglichkeiten der Evaluation präventiver Handlungsansätze und Netzwerke der Kinder-, Jugend und Gesundheitshilfe, Expertise zum 9. Kinder- und Jugendbericht der Landesregierung Nordrhein-Westfalen

Elkeles T (2006) Evaluation von Gesundheitsförderung und Evidenzbasierung? In: Bödeker W, Kreis J (Hrsg) Evidenzbasierte Gesundheitsförderung und Prävention. Wirtschaftsverlag NW, Bremerhaven, S 111–153

Elkeles T, Kirschner W (2003) Evaluation im Gesundheitswesen, Hochschule für Angewandte Wissenschaften Hamburg, Hamburg

Friese M, Walter M (2005) Evaluation der Frühberatungsstelle Bremen Hemelingen, Abschlussbericht 2005 www.soziales.bremen.de/sixcms/media.php/13/Evaluation%20Fruehberatung.pdf (Stand: 03.03.2012)

Lengning A (2010) Goldstandards für einen wissenschaftlichen Nachweis der Wirksamkeit und Effektivität einer Intervention im Bereich Früher Hilfen als Voraussetzung für ihre Verbreitung, Bundesgesundheitsbl – Gesundheitsforsch – Gesundheitsschutz 53: 1056–1060

Lengning A, Zimmermann P (2009) Expertise Interventions- und Präventionsmassnahmen im Bereich Früher Hilfen, Internationaler Forschungsstand, Evaluationsstandards und Empfehlungen für die Umsetzung in Deutschland (Hrsg): Nationales Zentrum Frühe Hilfen (NZFH)

Macintyre S (2011) Good intentions and received wisdom are not good enough: the need for controlled trials in public health, J Epidemiol Community Health 65: 564–567

MASF (o. Jahr) www.brandenburg.de/sixcms/media.php/4055/sozialindex_ogd2007.pdf (Stand: 18.09.2012)

MASF (2011) Forschung Beratung + Evaluation: Evaluation der Netzwerke Gesunde Kinder im Land Brandenburg im Jahr 2011 – Zusammenfassung der Ergebnisse www.masf.brandenburg.de/cms/detail.php/bb1.c.215996.de (Stand: 18.09.2012)

Pawson R, Tilley N (o. Jahr) Realist Evaluation Ray Pawson & Nick Tilley www.communitymatters.com.au/RE_chapter.pdf (Stand: 13.03.2012)

Pawson R, Tilley N (2009) Realistic Evaluation, Sage Publications, London, p.xiii

Reinisch S (2011) Entwicklung von Kindern in Beziehung, In: Thapa-Görder N, Voigt-Radloff S (Hrsg) Prävention und Gesundheitsförderung – Aufgaben der Ergotherapie, Georg Thieme Verlag, S 85–89

Rossi PH et al. (1992) Programmevaluation Einführung in die Methoden angewandter Sozialforschung, Ferdinand Enke Verlag, Stuttgart

Sass S (o. Jahr) Neuvola als Modell des aktivierenden Sozialstaates in Finnland library.fes.de/pdf-files/bueros/schwerin/04974.pdf (Stand: 02.03.2012)

Schmacke N (2009) Was bringt ein evidenzbasierter Ansatz in Prävention und Gesundheitsförderung? In: Kolip P, Müller VE (Hrsg) Qualität von Gesundheitsförderung und Prävention, Hans Huber

Schroder W (o. Jahr) www.masf.brandenburg.de/sixcms/media.php/4055/VorsorgenderSozialstaat.pdf (Stand: 18.09.2012)

Tschöpe-Scheffler S, Niermann J (2002) Evaluation des Elternkurskonzepts »Starke Eltern – Starke Kinder« des Deutschen Kinderschutzbundes, Forschungsbericht Fachhochschule Köln, Fakultät für Angewandte Sozialwissenschaft, Köln

Ziegler H (2010) Ist der experimentelle Goldstandard wirklich Gold wert für eine Evidenzbasierung der Praxis Früher Hilfen?. Bundesgesundheitsbl – Gesundheitsforsch – Gesundheitsschutz 53: 1061–1066

Evaluation der Fördertätigkeit des Fonds Gesundes Österreich

Karin Waldherr, Gerlinde Rohrauer-Näf, Monika Simek, Gudrun Braunegger-Kallinger, Christa Peinhaupt

Zusammenfassung

In dem Beitrag wird das Grobkonzept zur formativen Evaluation der Fördertätigkeit und der damit zusammenhängenden Aktivitäten des Capacity Building des Fonds Gesundes Österreich (FGÖ), der bundesweiten Koordinationsstelle für Gesundheitsförderung und Primärprävention, durch das Ludwig Boltzmann Institut Health Promotion Research (LBIHPR) vorgestellt. Beim vorliegenden Evaluationsgegenstand sind grundsätzlich zwei Ebenen zu unterscheiden: die untere Ebene besteht aus einzelnen Aktivitäten (geförderte Projekte, Aktivitäten zum Capacity Building), die auf einer übergeordneten Meta-Ebene insgesamt ein kohärentes Gesamtprogramm ergeben sollen. Ziel der Evaluation ist es, systematisches Lernen auf dieser den einzelnen Projekten und Aktivitäten übergeordneten Meta-Ebene zu ermöglichen. Nicht nur das Gesamtprogramm, sondern auch die einzelnen Aktivitäten selbst sind komplexe adaptive Systeme. In dem Beitrag werden daher zentrale Charakteristika und universelle Prinzipien komplexer adaptiver Systeme sowie Konsequenzen einer systemischen Sichtweise für die Evaluation dargestellt. Es wird argumentiert, dass ein nutzenorientierter Ansatz handlungsleitend sein sollte, bei dem die Evaluation bestmöglich an die Informationsbedürfnisse des Systems angepasst ist, um Lernen zu ermöglichen. Da die Ergebnisse des Gesamtsystems auf die Interaktionen der Systemkomponenten zurückzuführen sind, erweist sich die einfachste Form der Komplexitätsreduktion – die isolierte Betrachtung der einzelnen Systemkomponenten – als ungeeignet. Aus systemischer Sicht ist der Blick auf Wirkungszusammenhänge wesentlich, wobei eine Fokussierung auf relevante Schwerpunkte im System vorgeschlagen wird. Notwendig ist in diesem Zusammenhang die Explikation der Programmlogik. Um der Komplexität und Dynamik des Systems gerecht zu werden, ist für die Evaluation außerdem Flexibilität und Triangulation von Strategien, Methoden, Perspektiven und Zeitpunkten erforderlich.

1 Ausgangslage

Der Fonds Gesundes Österreich (FGÖ) ist die zentrale Förder- und Koordinationsstelle für Gesundheitsförderung und Primärprävention in Österreich. Laut österreichischem Gesundheitsförderungsgesetz (GfG 1998) umfassen die Aufgaben des FGÖ Information, Aufklärung und Öffentlichkeitsarbeit, Unterstützung wissenschaftlicher und praxisorientierter Projekte, Unterstützung beim Aufbau von Strukturen sowie Unterstützung bei Fortbildung und Vernetzung mit den globalen Zielen »*Erhaltung, Förderung und Verbesserung der Gesundheit der Bevölkerung im ganzheitlichen Sinn und in allen Phasen des Lebens*« und »*Aufklärung und Information über vermeidbare Krankheiten sowie über die Gesundheit beeinflussenden seelischen, geistigen und sozialen Faktoren*«.

Zur Umsetzung seiner Aufgaben steht dem FGÖ ein jährliches Budget von 7,25 Millionen Euro, ausgeschüttet aus öffentlichen Mitteln, zur Verfügung. Im Jahr 2006 wurde der FGÖ als Geschäftsbereich in die neu gegründete »Gesundheit Österreich GmbH (GÖG)« eingegliedert. Weitere Geschäftsbereiche der GÖG sind das »Österreichische Bundesinstitut für Gesundheitswesen (ÖBIG)« und seit 2007 das neu gegründete »Bundesinstitut für Qualität im Gesundheitswesen (BIQG)«. Durch die Schaffung der GÖG sollte eine Abstimmung von Strukturplanung, Gesundheitsförderung und Qualitätssicherung in Österreich erreicht werden.

2 Arbeitsprogramm[1] und Qualitätsentwicklungsmaßnahmen des FGÖ

Die bisherige Arbeit des FGÖ konzentrierte sich inhaltlich vorwiegend auf die Settings »Kindergarten und Schule«, »Betrieb« sowie »kommunales Setting« und die thematischen Schwerpunkte »Bewegung und Ernährung« sowie »Seelische

1 Wir definieren ein Programm als Gruppe verschiedener, untereinander koordinierter Maßnahmen, die der Erreichung gemeinsamer Ziele (Programmziele) dienen.

Gesundheit«. Die Umsetzung erfolgt in Form von Projektförderung sowie FGÖ-Initiativen und -Angeboten (Fort- und Weiterbildungsprogramm/Capacity Building, Information und Aufklärung, Unterstützung von Selbsthilfe, Vernetzungsaktivitäten, Beauftragung von Forschungs- und Evaluationsprojekten, Veranstaltungen). Förderungswürdige Projekte der Gesundheitsförderung und Primärprävention orientieren sich an einem umfassenden Gesundheitsbegriff, an den Prinzipien Nachhaltigkeit, gesundheitliche Chancengleichheit, Zielgruppenorientierung, Setting- und Determinantenorientierung, Ressourcenorientierung und Empowerment sowie Partizipation der Akteurinnen und Akteure des Settings und Qualitätskriterien, die internationalen Standards entsprechen (vgl. Fonds Gesundes Österreich 2012a, 19ff; Gesundheitsförderung Schweiz 2007; European Project Getting Evidence into Practice, NIGZ, VIG 2005; Jahn, Kolip 2002; Lehmann et al. 2006). Der FGÖ fördert Projekte von unterschiedlicher Dauer und Laufzeit in der Regel im Umfang von ein bis zwei Drittel der Gesamtprojektkosten. Im Jahr 2011 erhielten 125 Projekte eine Förderzusage (rund 27 % praxisorientierte Projekte im betrieblichen Setting, 21 % praxisorientierte Projekte im kommunalen Setting sowie 28 % praxisorientierte Projekte unterschiedlicher thematischer Ausrichtung, 22 % Fort- und Weiterbildung sowie Vernetzung, und einige wenige internationale Projekte) mit einem Gesamtfördervolumen von ca. 5 Millionen Euro (Fonds Gesundes Österreich 2012c). Im Arbeitsprogramm 2012 wird – im Einklang mit der WHO-Strategie »Gesundheit für alle« – gesundheitliche Chancengleichheit als handlungsleitendes Prinzip noch stärker in den Vordergrund gerückt. In Übereinstimmung damit werden Beratungs- und Sozialeinrichtungen als wichtige Settings genannt und die Notwendigkeit zielgruppenorientierter Zugänge unterstrichen (vgl. Fonds Gesundes Österreich 2012b). Evaluation und Qualitätsentwicklung sieht die FGÖ-Geschäftsstelle als wichtigen Bestandteil des Projektmanagements und als wesentlichen Beitrag zum kontinuierlichen Lernen und zur systematischen Weiterentwicklung von Gesundheitsförderungs- und Präventionsmaßnahmen (vgl. Fonds Gesundes Österreich 2012b, S. 10). Projekte mit einer Förderungssumme ab 72.000 Euro müssen extern evaluiert werden. Das zur Verfügung stehende Budget beträgt im Durchschnitt ungefähr 10 Prozent der anerkannten Gesamtprojektkosten. Kleinere Projekte können sich zwischen einer externen Evaluation oder einer Selbstevaluation entscheiden. Die Einholung entsprechender Angebote, die Definition der Fragestellungen für die Evaluation, die Entwicklung von Konzepten und die Umsetzung der Evaluation erfolgt jeweils spezifisch für die Einzelprojekte durch die Fördernehmer/-innen bzw. Evaluatorinnen/Evaluatoren selbst. Sie werden unterstützt durch Leitfäden und Checklisten bzw. durch das Feedback ihrer Ansprechperson im FGÖ.

Obwohl die routinemäßige Evaluation von Programmen oder Strategien im Allgemeinen noch nicht als vollständig etabliert gilt, steigt das Verständnis für deren Bedeutung: Einerseits können Programme und Verfahrensweisen dadurch verbessert und andererseits die Einhaltung von Verantwortlichkeiten überprüft werden (vgl. Kahan 2008, S. 5; Lorenzen 2011). Der FGÖ ließ bisher zwei externe Evaluationen durchführen (Kirschner et al. 2002, Kirschner et al. 2006). Beide Evaluationen bezogen sich auf einen Zeitraum von jeweils etwa drei Jahren (Dreijahresprogramme 1998–2001 und 2002–2005) und waren summativ ausgerichtet. Im Jahr 2008 ging der FGÖ einen Kooperationsvertrag mit dem damals neu gegründeten Ludwig Boltzmann Institut Health Promotion Research (LBIHPR) ein und setzte somit ein externes Expertinnen- und Expertenteam für evaluative Fragestellungen ein. In den ersten Jahren der Kooperation wurde die Qualität der vom FGÖ in den ersten zehn Jahren seines Bestehens geförderten Projekte und deren Evaluationen umfassender untersucht als dies im Rahmen der beiden externen Evaluationen möglich war und deren erzielte Ergebnisse sowie berichtete förderliche und hinderliche Faktoren zusammenfassend dargestellt. Für die nächsten Jahre der Kooperation sahen sowohl FGÖ als auch LBIHPR die Notwendigkeit einer formativ ausgerichteten, systematisch in Programmsteuerung und -management integrierten Evaluation der Fördertätigkeit und des Capacity Building als weiteren Entwicklungsschritt.

3 Zielsetzung der Evaluation

»Begründet ist die Durchführung einer formativen Programmevaluation des FGÖ mit der Notwendigkeit einer Weiterentwicklung der Aktivitäten und Initiativen des FGÖ unter Beachtung internationaler Entwicklungen und in Abstimmung mit den Aktivitäten anderer nationaler Akteurinnen und Akteure.« (Fonds Gesundes Österreich 2012b, S. 28). Hauptzweck und Anspruch der Evaluation ist es, systematisches Lernen auf einer den einzelnen Projekten und Maßnahmen übergeordneten Meta-Ebene als Grundlage für gezielte inhaltliche Weiterentwicklung zu ermöglichen. Die Evaluation soll als Informationsgrundlage dienen, um beispielsweise Charakteristika erfolgreicher Projekte, geeignete Strukturen für den Wissenstransfer sowie eventuell nicht gedeckten Bedarf in der Gesundheitsförderung und Primärprävention in Österreich zu identifizieren. Regelmäßige summative Zwischenbilanzierungen bezüglich erzielter Ergebnisse können neben der Nutzung als rationale Grundlage für Entscheidungen der formativen Programmsteuerung auch der Legitimation gegenüber Entscheidungsträgerinnen/Entscheidungsträgern, Politik und Öffentlichkeit dienen. Entsprechend einer Einteilung von Stockmann (2006) sind die übergeordneten miteinander verbundenen Ziele der Evaluation somit: Gewinnen von Erkenntnissen; Schaffen von Transparenz, um einen Dialog zu ermöglichen und Dokumentation des Erfolgs. Wie jede Evaluation übt sie damit implizit natürlich auch eine gewisse Kontrolle aus.

4 Der Evaluationsgegenstand

Beim vorliegenden Evaluationsgegenstand sind grundsätzlich zwei Ebenen zu unterscheiden: die untere Ebene besteht aus einzelnen Aktivitäten (geförderte Projekte, Aktivitäten zum Capacity Building), die auf einer übergeordneten Meta-Ebene insgesamt ein kohärentes Gesamtprogramm ergeben sollen. Nicht nur das Gesamtprogramm, sondern auch die einzelnen Aktivitäten selbst sind *komplexe adaptive Systeme*, welche folgendermaßen charakterisiert sind (für eine ausführlichere Darstellung vgl. z. B. Bar-Yam 2003; Holland 1992, 2006; Baecker 2009):

- Sie bestehen aus mehreren, miteinander in Wechselwirkung stehenden Elementen, die fähig sind in Form von Adaption auf die vielfältigen Einflüsse aus ihrer Umgebung zu reagieren, aus Erfahrung zu lernen und Konsequenzen ihrer Reaktionen auf die Umwelt zu antizipieren.
- Sie stehen zwar in Kontakt mit ihrer Umgebung, die gestaltenden und beschränkenden Einflüsse gehen jedoch von den Elementen des sich organisierenden Systems selbst aus (Selbstorganisation), sie produzieren sich prinzipiell selbst und schließen an ihre eigenen Strukturen an (Selbstreferenz). Das bedeutet, sie können nicht von außen hergestellt beziehungsweise kontrolliert werden.
- Wirkungszusammenhänge sind im Allgemeinen nichtlinear, d. h. die gleiche Maßnahme kann in verschiedenen Systemen zu vollkommen unterschiedlichen Ergebnissen führen.
- Auch wenn die Wechselwirkungen zwischen den Elementen lokal sind, sind Auswirkungen in der Regel global.
- Im Gegensatz zu lediglich komplizierten Systemen zeigen komplexe Systeme die Eigenschaft der »Emergenz« (= Herausbilden neuer Strukturen oder Eigenschaften des Systems aufgrund der Interaktionen zwischen den Elementen des Systems). Emergente Eigenschaften sind auf das Zusammenwirken der Elemente des Systems zurückzuführen und sind daher nicht durch getrennte Analyse der einzelnen Systemkomponenten erklärbar. (Um beispielsweise verstehen zu können, warum ein Mensch gehen kann, reicht es nicht aus, das Bein, den Rumpf und das Gehirn getrennt zu betrachten, sondern es ist notwendig die Interaktionen zwischen diesen zu verstehen.)
- Die einzelnen Elemente, wechselseitige Beeinflussungen im System sowie Probleme und Ziele verändern sich mit der Zeit.

Ziel der Evaluation ist es, Erkenntnisse über die emergenten Eigenschaften auf der globalen Ebene des Gesamtprogrammes zu erhalten. Da es hierfür essentiell ist die Interaktionen zwischen den einzelnen Akteurinnen und Akteuren zu verstehen, muss eine Evaluation des entstehenden Gesamtprogrammes mehr sein als die simple Addition der Evaluationen der einzelnen Maßnahmen und Akti-

vitäten (vgl. Holland 1992, 2006; Bar-Yam 2003; Kanfer et al. 2006). Es müssen sowohl Strukturen und Prozesse aller Aktivitäten als auch Vernetzungen und Interaktionen einzelner Aktivitäten und Bereiche berücksichtigt werden, also die einzelnen Aktivitäten im Gesamtkontext betrachtet werden.

Sowohl Gesamtprogramm als auch Einzelaktivitäten sind eingebettet in die Gesundheitsförderungslandschaft in Österreich, welche von der GÖG, aber auch anderen wichtigen Akteurinnen und Akteuren (beispielsweise regionale Gesundheitsförderungseinrichtungen oder Gebietskrankenkassen) konstituiert wird. Jedes System ist somit eingebunden in ein hoch komplexes Netzwerk von Einflussgrößen, die sich wiederum gegenseitig beeinflussen, was ebenfalls zu berücksichtigen ist.

Da es unrealistisch ist, dass eine Evaluation alle Einflussgrößen und ihr wechselseitiges Zusammenwirken analysieren oder kontrollieren kann, ist eine geeignete Form der Komplexitätsreduktion notwendig, um handlungsfähig zu bleiben. Ein interdisziplinärer systemischer Ansatz beruhend auf der Kybernetik (vgl. z. B. Kaufmann 2007), der Systemtheorie (vgl. z. B. Baecker 2009) und der Komplexitätstheorie (z. B. Bar-Yam 2003) bietet die notwendige theoretische Basis. *"... for evaluation of complex social systems to be effective, the evaluation process must take into account the theoretical understanding of complex systems."* (McDonald, Kay 2006, S. 2). Da komplexe adaptive Systeme – seien es nun biologische, technische oder soziale Systeme – einige universelle Eigenschaften aufweisen, können Konzepte und Tools aus verschiedensten Forschungsbereichen, die mit komplexen adaptiven Systemen konfrontiert sind, hilfreich sein. Dies sind beispielsweise Sozial- und Wirtschaftswissenschaften (insbesondere Qualitätsmanagement, Organisationsentwicklung/Lernende Organisationen, z. B. Senge 1990; Dooley 1997), systemische Therapie (z. B. Selbstmanagementtherapie, Kanfer et al. 2006), Kommunikationstheorie (Shannon, Weaver 1949/1963), Informationswissenschaften und Wissensmanagement (vgl. z. B. Shaxson 2009), Naturwissenschaften und Technik – insbesondere Neurobiologie, Neuronale Netze und Künstliche Intelligenz (z. B. Entwicklung von »organischen« Computersystemen bzw. lernenden Systemen, vgl. z. B. Müller-Schloer et al. 2004) – sowie Computerwissenschaften (z. B. Evolutionäres Programmieren bzw. Genetische Algorithmen, vgl. z. B. Holland 1992, 2006). Gemeinsame Basis sind die drei kybernetischen Elemente Regelung (Feedback), Information (Kommunikation) und Wirkungszusammenhänge (Komplexität, vgl. Kaufmann 2007).

5 Evaluationsansatz

5.1 Grundsätzliche Überlegungen

5.1.1. Nutzenorientierung

Das LBIHPR orientiert sich in seiner Arbeit an den Standards der Deutschen Gesellschaft für Evaluation (DeGEval 2008). Aus einer systemischen Sichtweise ist vor allem die in den Standards geforderte Eigenschaft der »Nützlichkeit« einer Evaluation hervorzuheben: Wenn wir das Gesamtprogramm als komplexes adaptives System begreifen, folgt daraus, dass eine Leistungssteigerung vor allem durch Lernen des Systems aufgrund von Erfahrungen sowie Antizipieren von Reaktionen der Umwelt möglich ist (vgl. auch Connick, Innes 2001). Die Evaluation muss bestmöglich an die Informationsbedürfnisse des Systems angepasst sein, oder mit anderen Worten ausgedrückt, das System dort abholen, wo es steht, um an die systemeigenen Strukturen anschlussfähig zu sein und das System nicht aus dem Gleichgewicht zu bringen. Als handlungsleitendes Prinzip liegt der Evaluation daher der Grundsatz der Nutzenorientierung zugrunde (vgl. Utilization-focused Evaluation; Patton 2008). Patton definiert eine nutzungsorientierte Evaluation als *"evaluation done for and with specific intended primary users for specific, intended uses"* (S. 37). Und weiter: *"As an evaluation unfolds, evaluators and primary intended users must work together to identify the evaluation that best fits their information needs and the programme's context and situation. This means negotiating with stakeholders, especially primary intended users and other key stakeholders, and adapting the evaluation design to financial, political, timing and methodological constraints and opportunities"* (Patton 2008). Für das vorliegende Evaluationsprojekt gilt: Die FGÖ-Geschäftsstelle ist primäre Nutzerin der Evaluationsergebnisse, die gezielte inhaltliche Weiterentwicklung der Förderschiene und des Capacity

Building für Gesundheitsförderung und Primärprävention ist deren spezifischer Nutzen. Die Evaluation muss also eng mit den Mitarbeiterinnen und Mitarbeitern der FGÖ-Geschäftsstelle zusammenarbeiten und systematisch in organisationsinterne Lernprozesse integriert werden. Da die Leistung des Gesamtsystems wesentlich von der Umsetzung in den geförderten Projekten und den Maßnahmen zum Capacity Building abhängt, ist die Partizipation der Förder- und Auftragnehmer/-innen erforderlich.

5.1.2 Blick auf Wirkungszusammenhänge

Komplexe adaptive Systeme sind dadurch charakterisiert, dass Wechselwirkungen zwischen ihren Komponenten zwar lokal, deren Auswirkungen aber zumeist global sind, und es daher »kritische Einflussgrößen mit großer Hebelwirkung« gibt. Eine Möglichkeit der Komplexitätsreduktion ist daher eine Fokussierung entsprechend der von Kanfer et al. (2006) als »Zoom-Objektiv-Metapher« bezeichneten Devise »global denken, lokal handeln« (Hendersen 1988). Ähnlich wie beim Fotografieren werden mit einem »Zoom-Objektiv« relevante Schwerpunkte von Systemen sehr detailliert betrachtet. In einer »Weitwinkeleinstellung« wird jedoch der Stellenwert dieser fokussierten Schwerpunkte im System berücksichtigt (Kanfer et al. 2006, S. 22). *"While studying the parts in isolation does not work, the nature of complex systems can be probed by investigating how changes in one part affect the others, and the behavior of the whole."* (Bar-Yam 2000/2005). Es geht also um die Suche nach Einflussgrößen/Faktoren auf lokaler Ebene, mit denen möglichst große globale Wirkungen zu erzielen sind.

5.1.3 Programmlogik

Selbstverständlich gibt es eine Vielzahl potenziell relevanter Faktoren. Um die Menge an Information erfassen zu können, implizit vorhandenes Wissen explizit zu machen sowie Annahmen über Wirkungszusammenhänge offenzulegen und überprüfbar zu machen, ist die Rekonstruktion der dem Programm zugrundeliegenden Ziele und Annahmen über Wirkungszusammenhänge unter Berücksichtigung des Programmkontextes hilfreich und notwendig (vgl. hierzu z. B. auch die ZiWi-Methode; von Unger et al. 2008). Daher wird in einem gemeinsamen Prozess mit den primären Nutzerinnen und Nutzern reflektiert, welche Ziele mit dem Förderprogramm des FGÖ erreicht werden sollen und welche Annahmen der Fördertätigkeit sowie den Aktivitäten zum Capacity Building für Gesundheitsförderung und Primärprävention zugrunde liegen. Das Modell folgt in der Grundlogik dem Outcome Model for Health Promotion (Nutbeam, 1998): i) Welche (emergenten) Strukturen und Eigenschaften soll das Gesamtprogramm aufweisen (= Outcomes), um einen bestimmten Impact zu erzielen? ii) »Welche Schritte/Maßnahmen müssen gesetzt werden (= Einflussgrößen mit potenziell großer Hebelwirkung), um bestimmte unmittelbare Outputs, die direkt vom FGÖ beeinflussbar sind, zu erzeugen?«, iii) »Welche anderen Einflussfaktoren könnten eine Rolle spielen?« Dieses Modell bietet die Grundlage zur Definition von Meilensteinen sowie relevanter Fragestellungen, Dimensionen und Indikatoren für die Evaluation.

5.1.4 Berücksichtigung von Komplexität und Dynamik

Um jene »Unterschiede zu finden, die Unterschiede machen« (vgl. Bateson 1972, S. 453), muss die Evaluation der Komplexität und Dynamik des Systems gerecht werden. Dies erfordert Flexibilität und Triangulation von Methoden, Perspektiven und Zeitpunkten (vgl. Eoyang, Berkas 1999). Feedback- und Feedforward-Schleifen dienen der periodischen Reflexion. Ähnlich der Vorgangsweise bei Gedankenexperimenten können verschiedene »Szenarien« gedanklich durchgespielt werden, also Vermutungen aufgestellt werden wie sich bestimmte Veränderungen auf das Gesamtsystem auswirken und welche Faktoren »störend« wirken könnten (Feedforward-Schleifen). Dies dient auch der Identifizierung von bestehenden Unsicherheiten und der Antizipation von »Überraschungen«. Meilensteinplanungen und regelmäßiges Feedback an alle Stakeholder sind wichtige Grundlagen für kontinuierliche Weiterentwicklung. Auch ständige, nicht vorhersagbare Veränderungen im System müssen beobachtet und beschrieben werden. »Schnappschüsse« zu festgelegten Beginn- und Endpunkten reichen daher nicht aus (vgl. Eoyang, Berkas 1999).

5.1.5 Micro- und Macro-Level

Einzelne Subsysteme eines komplexen adaptiven Systems erzeugen verschiedene »Variationen«, wobei manche Muster in gleicher Form in verschiedenen Subsystemen ablaufen. Die einzelnen Projekte stellen also eine Art »Simulationen« dar und können als Proxies dienen, die im Rahmen von Fallstudien beobachtet und verglichen werden können. Regelmäßige Feedback-Schleifen zur Beantwortung von Fragen wie beispielsweise »Welche Variationen wurden produziert?« (z. B. welche methodischen Zugänge wurden gewählt), »Welche Widerstände haben sich entwickelt und was waren die Gründe dafür?«, »Welche intendierten und nicht intendierten Effekte sind aufgetreten?« oder »Welche kritischen Ereignisse mit großen Auswirkungen auf das Gesamtsystem sind aufgetreten?« verbessern das Verständnis für die Ergebnisse dieser Prozesse in ihrem jeweiligen Kontext und helfen dabei, Gemeinsamkeiten und allgemeine Prinzipien in verschiedenen Subsystemen (z. B. verschiedenen Settings) sowie Spezifika zu erkennen.

5.2 Evaluationsdesign

Das Evaluationsdesign muss also der Komplexität und Dynamik des Systems entsprechen und daher die Triangulation von Strategien, Methoden, Perspektiven und Zeitpunkten vorsehen. Auf Basis dieser grundsätzlichen Überlegungen resultierte das von den Autorinnen erarbeitete iterative Evaluationsdesign bestehend aus:

1) Reflexion/Fokussierung: Reflexion und Konkretisierung der Ziele der Fördertätigkeit und des Capacity Building, Explikation der Programmlogik und Konkretisierung notwendiger Maßnahmen, Meilensteine und Indikatoren für deren Erreichung sowie Einigung bezüglich der Grenzen der Evaluation, der konkret zu beantwortenden Fragestellungen und des konkreten Evaluationsdesigns.
2) Formative Prozessevaluation: Die Koordinations- und Unterstützungsprozesse des FGÖ (z. B. interne Abstimmungsprozesse, Vorgaben für die Projektevaluationen, Erreichen von Multiplikatoren/Multiplikatorinnen durch Vernetzungsaktivitäten) werden als Faktoren mit potenziell großen globalen Wirkungen betrachtet und werden daher im Fokus der Prozessevaluation stehen. Da die Ergebnisse auf Ebene des Gesamtprogramms wesentlich durch die geförderten Projekte mitbestimmt werden, ist des Weiteren eine Verlinkung der verschiedenen Ebenen notwendig, d. h. es müssen auch Projektprozesse berücksichtigt werden (z. B. von den geförderten Projekten ausgehende Vernetzungsimpulse, Umsetzung der Vorgaben für Projektevaluationen in der Praxis, Erreichen von Multiplikatoren/Multiplikatorinnen in den Projekten,...).
3) Summative Zwischenbilanzen (Feedback-Schleifen): Diese sollten verschiedene Perspektiven berücksichtigen und dienen zur Feststellung, ob die Meilensteine erreicht wurden und inwieweit durch das Konglomerat der geförderten Projekte und der Capacity Building-Aktivitäten ein kohärentes Programm entstehen kann und die Programmziele erreicht werden können. Sie stellen wiederum Grundlage zur periodischen Reflexion und zur Planung der nächsten Meilensteine dar und bilden somit auch den Ausgangspunkt für die nächste Iteration mit einer aufgrund der Dynamik des Systems gegebenenfalls notwendigen Anpassung der Programmlogik und des Evaluationskonzeptes. Somit sind die summativen Zwischenbilanzen integrativer Bestandteil der formativen Evaluation. Des Weiteren dienen sie aber auch der Dokumentation erzielter Ergebnisse.

Das Design sieht einerseits ein standardisiertes Vorgehen vor, um Vergleichbarkeit der Ergebnisse herzustellen und systematisches Lernen zu ermöglichen. Dies beinhaltet einige einfache Richtlinien sowie ein projektübergreifendes Indikatorensystem für Projektevaluationen und -dokumentationen. Die spezifische Herausforderung dabei ist, ein Indikatorensystem zu erstellen, das einerseits ausreichend Informationen liefert, andererseits aber die Projektnehmer/-innen nicht überfordert. Andererseits sind auch Fallstudien erforderlich, um Prozesse und Wirkungszusammenhänge in ihrem jeweiligen Kontext zu beobachten und besser verstehen zu können.

6 Zwischenergebnisse

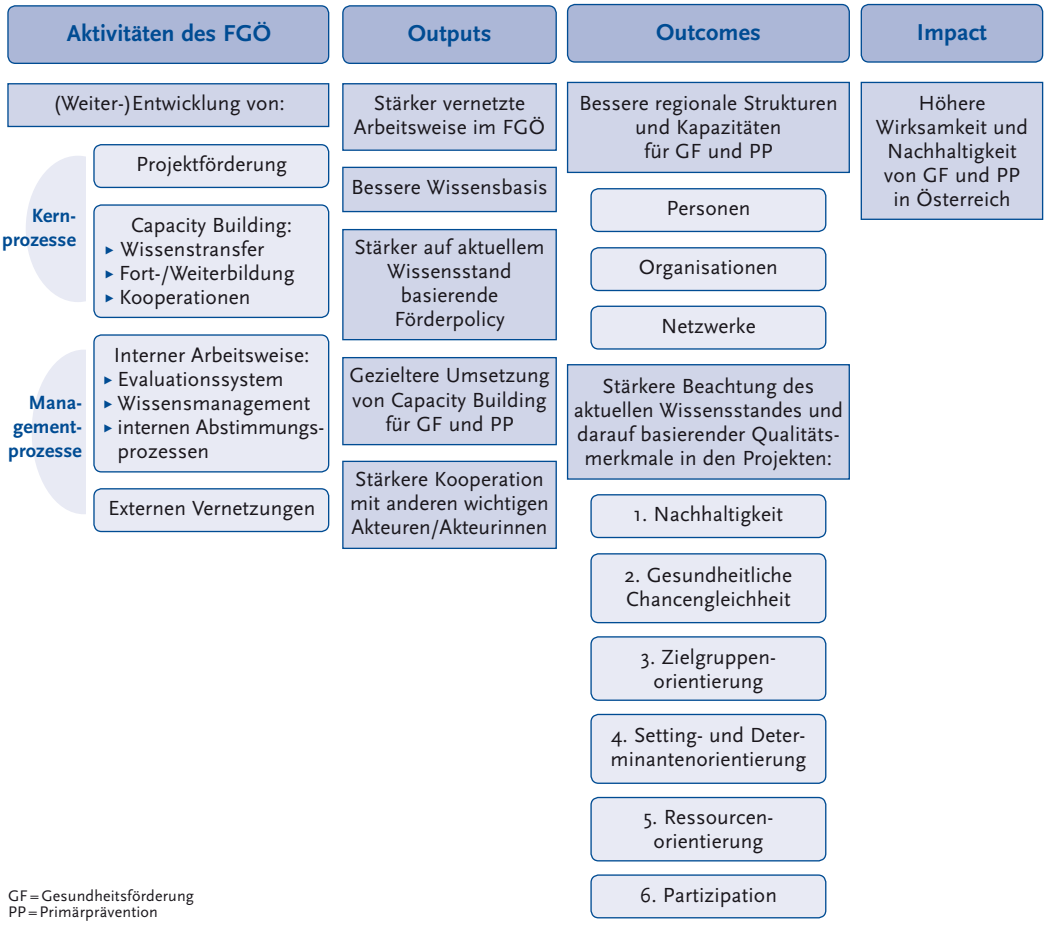
Die Evaluation startete im Jänner 2012. Bis Juni 2012 wurden drei Workshops mit den inhaltlich verantwortlichen Mitarbeiterinnen und Mitarbeitern des FGÖ abgehalten. Die Schwerpunkte waren zunächst Erzeugung von "Readiness for Evaluation" sowie Reflexion der Ziele der Fördertätigkeit und zugrundeliegender Wirkungsannahmen (Ausgangsversion der Programmlogik, s. Abbildung 1):

Grundannahme ist, sehr vereinfacht dargestellt, dass durch inhaltlich strategische Weiterentwicklung der Förderschiene bei gleichzeitig stärkerer Systematisierung des Capacity Buildung für Gesundheitsförderung und Primärprävention aktuelles Wissen jene Personen erreicht, die tatsächlich in der Praxis der Gesundheitsförderung und Primärprävention tätig sind, wodurch

- die Qualitätsmerkmale der Gesundheitsförderung und Primärprävention stärkere Beachtung in der Praxis finden und
- bessere lokale/regionale Strukturen und Kapazitäten für Gesundheitsförderung und Primärprävention auf individueller, organisationaler und organisationsübergreifender Ebene entstehen,

was schließlich in höherer Wirksamkeit und Nachhaltigkeit der Gesundheitsförderung und Primärprävention in Österreich resultieren sollte.

Abbildung 1
Programmlogik



GF = Gesundheitsförderung
PP = Primärprävention

Darauf aufbauend erfolgte in einem weiteren partizipativen Prozess eine Konkretisierung der notwendigen Maßnahmen. Als Basis für die Weiterentwicklung der Kernprozesse des FGÖ – Projektförderung und Capacity Building für Gesundheitsförderung und Primärprävention – werden kontinuierliches, systematisches Lernen, eine stark vernetzte interne Arbeitsweise sowie externe Vernetzungen gesehen. Um die Förderpolicy und die Umsetzung in den Projekten auf Basis von Lernerfahrungen und Evidenz systematisch weiterentwickeln zu können, sind die Organisation des Wissens und der Wissenstransfer von zentraler Bedeutung. Das heißt, dass vorhandenes Wissen aus den geförderten Projekten und der Forschung systematisch strukturiert werden muss und vorhandene Strukturen zum Wissenstransfer effizient genützt werden müssen. Aufgrund der strukturierten Wissensbasis wird des Weiteren erkennbar, welches zusätzliche Wissen benötigt wird, und daher anhand der Projektdokumentationen und -evaluationen in den verschiedenen Settings und thematischen Schwerpunkten zukünftig generiert werden sollte. Auf dieser Basis kann ein Evaluationssystem in Form von Richtlinien für die Dokumentationen und Evaluationen inklusive eines übergreifenden Indikatorensystems erarbeitet werden (beziehungsweise zu späteren Zeitpunkten gegebenenfalls adaptiert werden). Durch diese Strukturierung wird Kohärenz zwischen den Settings und Themenschwerpunkten erzeugt, wird die strategische Übersicht erhöht, werden die Projektevaluationen bestmöglich an die Informationsbedürfnisse des Gesamtsystems angepasst und wird somit das Erkennen von Gemeinsamkeiten, universellen Prinzipien und bedeutenden Faktoren ermöglicht. Nicht zuletzt wird damit der enge funktionale Zusammenhang zwischen der Weiterentwicklung der Förderpolicy und dem in der Praxis erzeugten Wissen deutlich gemacht (vgl. Shaxson 2009).

7 Fazit

Entscheidend für Evaluationen von komplexen Programmen in der Gesundheitsförderung und Prävention ist eine interdisziplinäre systemische Sichtweise. Neben wichtigen theoretischen Grundlagen aus der Kybernetik, der Systemtheorie und der Komplexitätstheorie erweisen sich Konzepte, Erfahrungen und Tools aus anderen Forschungsbereichen, die ebenfalls mit komplexen adaptiven Systemen und den damit zusammenhängenden Problemen konfrontiert sind, als hilfreich.

Aus einer systemischen Sichtweise folgt, dass die Evaluation bestmöglich an die Informationsbedürfnisse des Systems angepasst sein muss, um Lernen zu unterstützen. Der nutzungsorientierte Ansatz erfordert eine enge Kooperation mit den primären Nutzerinnen und Nutzern, aber auch anderen wichtigen Stakeholdern, da eine Leistungssteigerung nur durch das Zusammenwirken aller Beteiligten möglich ist. Die Bereitschaft, die notwendige Zeit und Energie in die Evaluation zu investieren korrespondiert mit dem Nutzen, den diese nicht nur langfristig, sondern auch unmittelbar bringt. Gerade die enge Kooperation führt jedoch sehr rasch zu *sichtbarem*, von den primären Nutzerinnen und Nutzern wahrgenommenem, unmittelbarem Nutzen. Im vorliegenden Fall zeigt sich dies in einigen Aspekten.

In Zeiten von knappen Ressourcen sind vertiefte Prozesse der Reflexion, Fokussierung und Strukturierung oft nur begrenzt möglich. Durch den gewählten Evaluationsansatz werden gut vorbereitete (Zeit)Räume für solche Aufgaben geschaffen, die dem Austausch und der Kooperation der verschiedenen thematischen Arbeitsbereiche im FGÖ dienen. Wichtig ist hierbei, dass sich diese Evaluationsprozesse in die Kernprozesse des FGÖ optimal integrieren lassen. Evaluationsworkshops können mit internen Planungstreffen verbunden werden. So kann beispielsweise die Reflexion über Ziele und vermutete Wirkungsverläufe, die einige Sitzungen benötigte und sehr zeitaufwändig war, unmittelbar in die anstehende Weiterentwicklung des nächsten Arbeitsprogramms integriert werden. Darüber hinaus wurden durch die Evaluation die – implizit vorhandenen – Themen proaktive Steuerung von Evaluations- und Forschungsthemen zur Erzeugung relevanten Wissens, bessere Strukturierung und Koordination von internen Prozessen und Wissensmanagement explizit gemacht und in den Fokus gerückt.

Nicht nur für die primären Nutzer/-innen und andere wichtige Stakeholder, sondern auch für die Effektivität der Evaluationsforschung in der Gesundheitsförderung und Prävention sind die enge Kooperation und die daraus resultierenden

Synergien in einem transdisziplinären Ansatz bedeutsam: *"To understand what is information for a client, one must understand the client's task. To maximize the form, format, and schedule* (of an evaluation; ergänzt durch die Autorinnen des vorliegenden Beitrages), *one must understand not only the task, but also the client's psychology."* (Cohen 1999). Umso intensiver die Kooperation verläuft, desto rascher entwickeln Evaluatorinnen/Evaluatoren und Nutzer/-innen ein gemeinsames Verständnis über die zentralen Informationsbedürfnisse und erfolgt eine Ko-Evolution des Evaluationssystems. Schließlich ist auch eine Evaluation ein komplexes adaptives System, dessen Ergebnisse auf das Zusammenwirken aller Beteiligten zurückzuführen sind.

Literatur

Baecker D (Hrsg) (2009) Niklas Luhmann. Einführung in die Systemtheorie, 5. Auflage. Carl-Auer Verlag, Heidelberg

Bar-Yam Y (2000/2005) Significant points in the study of complex systems. New England Complex Systems Institute http://necsi.edu/projects/yaneer/points.html (Stand: 19.04.2012)

Bar-Yam Y (2003) Dynamics of Complex Systems. (Studies in Nonlinearity) Westview Press

Bateson G (1972) Steps to an Ecology of Mind, Reprint. Chicago: Chicago UP, 2000. (Zitiert nach Baecker D (2010), Die Texte der Systemtheorie) www.dirkbaecker.com/Texte.pdf (Stand: 12.04.2012)

Cohen EB (1999) Reconceptualizing Information Systems as a Field of the Transdiscipline Informing Science: From Ugly Duckling to Swan. Journal of Computing and Information Technology 7(3): 213–219

Connick S, Innes JE (2001) Outcomes of Collaborative Water Policy Making: Applying Complexity Thinking to Evaluation. IURD Working Paper Series, Institute of Urban and Regional Development, UC Berkeley http://escholarship.org/uc/item/03f3b4z9 (Stand: 12.04.2012)

Deutsche Gesellschaft für Evaluation (Hrsg) (2008) Standards für Evaluation (DeGEval-Standards). 4. Auflage. Mainz

Dooley KJ (1997) A Complex Adaptive Systems Model of Organization Change. Nonlinear Dynamics, Psychology, and Life Sciences 1 (1): 69–97

Eoyang G, Berkas T (1999) Evaluating performance in a complex adaptive system. In: Lissack M, Gunz H (Eds) Managing complexity in organizations. Westport, Connecticut: Quorum Books

European Project Getting Evidence into Practice, NIGZ, VIG (2005) European Quality Instrument for Health Promotion (EQUIHP) www.nigz.nl/gettingevidence (Stand: 12.04.2012)

Fonds Gesundes Österreich (Hrsg) (2012a) Leitfaden für Antragsteller und Antragstellerinnen und Fördernehmer und Fördernehmerinnen. Fonds Gesundes Österreich, Wien

Fonds Gesundes Österreich (Hrsg) (2012b) Arbeitsprogramm 2012. Fonds Gesundes Österreich, Wien

Fonds Gesundes Österreich (Hrsg) (2012c) Jahresbericht 2011, Geschäftsbereich Fonds Gesundes Österreich der Gesundheit Österreich GmbH. Fonds Gesundes Österreich, Wien

Gesundheitsförderungsgesetz – GfG (1998) Bundesgesetz über Maßnahmen und Initiativen zur Gesundheitsförderung, -aufklärung und -information. Bundesgesetzblatt für die Republik Österreich. Österreichische Staatsdruckerei, Wien, ausgegeben am 27. März 1998

Gesundheitsförderung Schweiz (Hrsg) (2007) Qualitätskriterien für Projekte, Version 5.0. quint-essenz: Qualitätsentwicklung in Gesundheitsförderung und Prävention, Gesundheitsförderung Schweiz www.quint-essenz.ch (Stand: 30.11.2007)

Hendersen H (1988) Gobal denken, lokal handeln. Politik und Ethik im Solarzeitalter. In: Lutz R (Hrsg) Pläne für eine menschliche Zukunft. Beltz, Weinheim, S 268–282

Holland JH (1992) Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control and Artificial Intelligence (Complex Adaptive Systems). MIT Press, Massachusetts

Holland JH (2006) Studying complex adaptive systems. Jrl Syst Sci & Complexity 19: 1–8

Jahn I, Kolip P (2002) Die Kategorie Geschlecht als Kriterium für die Projektförderung der Gesundheitsförderung Schweiz. Bremer Institut für Präventionsforschung und Sozialmedizin (BIPS), Bremen www.bips.uni-bremen.de/data/jahn_gesundheitsfoerderung_2002.pdf (Stand: 12.04.2012)

Kahan B (2008) Excerpts from Review of Evaluation Frameworks. Saskatchewan Ministry of Education www.idmbestpractices.ca/pdf/evaluation-frameworks-review.pdf (Stand: 11.05.2010)

Kanfer FH, Reinecker H, Schmelzer D (2006) Selbstmanagement-Therapie, 5. Auflage. Springer, Berlin

Kaufmann M (2007) Der Baum der Kybernetik. Die Entwicklungslinien der Kybermetik von den historischen Grundlagen bis zu ihren aktuellen Ausformungen. proEval Verlag, Dornbirn

Kirschner R, Elkeles T, Kirschner W (2002) Evaluation der Tätigkeit des Fonds Gesundes Österreich 1998–2001. Ergebnisbericht. Fonds Gesundes Österreich, Wien

Kirschner W, Kirschner R, Lenk S et al. (2006) Evaluation der Tätigkeit des Fonds Gesundes Österreich im Programmzeitraum 2002–2005. Zusammenfassung der Ergebnisse. Fonds Gesundes Österreich, Wien

Lehmann F, Geene R, Kaba-Schönstein L et al. (2006) Kriterien guter Praxis in der Gesundheitsförderung bei sozial Benachteiligten. Ansatz – Beispiele – Weiterführende Informationen. BzgA, Köln www.bzga.de/pdf.php?id=df0211e4054429896a7fe5568c7c7e99 (Stand: 12.04.2012)

Lorenzen HP (2011) Möglichkeiten und Grenzen der Evaluierbarkeit komplexer Interventionen. Einführungsvortrag zum Workshop »Evaluation komplexer Interventionsprogramme in der Prävention: Lernende Systeme, lehrreiche Systeme?«. Berlin, 05.12.2011

McDonald DM, Kay N (2006) Towards an evaluation framework for complex social systems. Proceedings of the Sixth International Conference on Complex Systems, June 25-30, Boston, USA
www.necsi.edu/events/iccs6/viewpaper.php?id=221 (Stand: 12.04.2012)
Müller-Schloer C, v d Malsburg C, Würtz RP (2004) Organic Computing. Informatik_Spektrum_4_August_2004, 2–6
Nutbeam D (1998) Evaluating health promotion – progress, problems and solutions. Health Promotion International, 13: 27–44
Patton MQ (2008) Utilization-focused evaluation, 4th ed. Sage, California
Senge P (1990) The fifth discipline. Doubleday, New York
Shannon CE, Weaver W (1949/1963) The Mathematical Theory of Communication. University of Illinois Press
Shaxson L (2009) Structuring policy problems for plastics, the environment and human health: reflections from the UK. Phil Trans R Soc B 364: 2141–2151
Stockmann R (2006) Evaluation und Qualitätsentwicklung. Eine Grundlage für wirkungsorientiertes Qualitätsmanagement. Waxmann, Münster
von Unger H, Block M, Wright MT (2008) Partizipative Qualitätsentwicklung – Methodenkoffer – Entwicklung lokaler Ziele und Wirkungswege (ZiWi-Methode)
www.partizipative-qualitaetsentwicklung.de/subnavi/methodenkoffer/ziwi-methode.html (Stand: 12.04.2012)

Chaos ist keine gute Idee: Von der Petitio Principii zu definitorischer Klarheit. Eine Nachbetrachtung

Manfred Wildner

Wissenschaftliche Diskussionen sind, auch wenn sie bisweilen emotional geführt werden, in ihrer Essenz durch den Austausch von empirischen Befunden und logischen Schlüssen geprägt. Die logische Figur einer *Petitio Principii* beispielsweise ist dadurch gekennzeichnet, dass die Schlussfolgerung eines Argumentes bereits in den eingeführten Prämissen versteckt ist. Dieser Sachverhalt kann als Schwäche einer Argumentation ausgelegt werden. Gleichzeitig kann im positiven Sinn festgestellt werden, dass ein solcher Schluss logisch gültig ist.

Welcher Zusammenhang besteht mit der Diskussion um die Evaluation komplexer Interventionen? Mit Bezug zu diesem Feld wurde kritisch angemerkt, dass Komplexität vielleicht gar keine so gute Idee sei, nämlich dann, wenn damit Eigenschaften wie Unvorhersagbarkeit, multiple Interdependenzen, das Auftreten unvorhergesehener Effekte und unerwartetes Chaos verbunden sind (Wolfgang Bödeker). In der Theorie komplexer Systeme werden genau solche Eigenschaften als systemimmanent aufgeführt: Unerwartete Wechselwirkungen zwischen den einzelnen Teilen, Nicht-Linearität, fehlende Vorhersagbarkeit der systemischen Reaktionen und anderes mehr. Klassische naturwissenschaftliche Beispiele für komplexe, nicht-lineare Prozesse sind das Wetter und der Herzrhythmus, Beispiele in den Geistes- und Sozialwissenschaften sind Börsenkurse und Konjunkturentwicklungen. Zugespitzt kann auf Basis dieser definitorischen Grundlage zu Recht die Frage gestellt werden, ob komplexe Interventionen in der Prävention überhaupt ethisch zulässig sind: Wenn nämlich bei solchen Interventionen wissentlich unvorhersagbare Effekte in Kauf genommen werden müssen.

Überlegungen, die zum Nachdenken anregen. Zum Einen stellt sich die Frage, ob der Begriff »komplexe Intervention« im Bereich von Prävention und Gesundheitsförderung tatsächlich im Rückgriff auf die systemtheoretisch definierte Komplexität mit Chaos, Turbulenzen, fehlender Reproduzierbarkeit und schwieriger Vorhersehbarkeit verknüpft werden sollte. Einzelne Ergebnisse der Evaluationsforschung deuten dabei durchaus in diese Richtung (Thomas Kliche).

Ein zweiter möglicher Umgang mit diesem augenscheinlichen Dilemma wäre eine differenzierte Herangehensweise, beispielsweise eine Unterscheidung von *komplexen* Mehrebenen- oder Multikomponenten-*Interventionen*, Interventionen in *komplexen Umgebungen* und *komplexen Perspektiven* beispielsweise aus unterschiedlichen Systemebenen auf Problemstellungen, Interventionseffekte und Evaluationsansätze. Bezüglich schwer zu definierender (und zu lösender) komplexer Probleme wurde der Begriff der *"wicked problems"* in der Sozialplanung geprägt (Rittel, Webber 1973). Dass eine einfache Intervention in komplexen bzw. instabilen dynamischen Systemen unvorhersagbare Auswirkungen haben kann, wurde in der bekannten Frage formuliert *"Does the flap of a butterfly's wings in Brazil set off a tornado in Texas?"* (Lorenz 1993; Lorenz 1963).

Ein Argument, eine Vorhersagbarkeit der Effekte von Interventionen gar nicht erst zu versuchen? Die Komplexität einschränkend ist zu sagen, dass emergente Eigenschaften in einem System in der Regel dimensional geführte Eigenschaften sind, d. h. beim Wetter treten unerwartete Wetterereignisse auf, aber keine Erdbeben, bei der Prävention treten auch nicht Ereignisse völlig anderer Natur auf. Um die Analogie des Wetters aufzugreifen: In Anerkennung der Tatsache, dass das Phänomen Wetter ein komplexer atmosphärischer Prozess mit begrenzter Vorhersagbarkeit ist, ist es gleichzeitig gelungen, mit verbesserten Methoden, insbesondere durch Computersimulationen, mehrtägige valide Vorhersagen zu erstellen. Auch wenn es weiterhin schwierig ist, »Wetter zu machen«, also ein gewünschtes Verhalten dieses komplexen Systems flächendeckend nachhaltig und planbar zu erzielen, sind erfolgreiche *Modellierungen von* und daraus abgeleitete *Interventionen in* einem solchen System durchaus möglich. In der vorgestellten Analogie könnte bei vorhergesagtem Regenwetter als präventive Intervention z. B. empfohlen werden,

einen Regenschirm mitzunehmen. Bei der Vorhersage von Hitzeperioden könnten umfassende Empfehlungen mit multiplen Teilkomponenten gegeben werden: An Krankenhäuser, Altenheime, Ärzte und Zuhause pflegenden Personen bezüglich Kleidung, Flüssigkeits- und Nahrungsaufnahme, Medikamentengabe, Beeinflussung des Raumklimas u. a. m. Man beachte: Bestandteil dieser einfachen bzw. Multikomponenten-Intervention war eine erfolgreiche, ausreichend quantitative Modellierung, ein Impuls, der auch für die Gesundheitswissenschaften gegeben wurde (West 2012).

Das Komplexitätsproblem könnte zum Dritten auch dadurch aufgegriffen werden, dass zur Klärung der Rede über Sachverhalte, welche gerne mit den Worten »komplex« bezeichnet werden, definitorische Einigungen getroffen werden. Mögliche alternative oder komplementäre Bezeichnungen bei Interventionen wären »Mehrebenen-Intervention«, »Intervention mit multiplen Komponenten«, »komplizierte Intervention« – wobei der Begriff kompliziert im medizinischen Bereich die Assoziation zu unerwünschten Nebenwirkungen in sich trägt –, »konzertierte Aktionen/Aktionspläne«, »multiple Programmbausteine« oder anderes (Alf Trojan).

Damit ist eine Diskussion nur angestoßen, noch lange nicht geführt. Wohin die naheliegenden Thesen »*komplexe Probleme bedürfen (angemessen) komplexer Interventionen*« und »*komplexe Interventionen bedürfen (angemessen) komplexer Evaluationen*« noch führen werden, darf mit Spannung erwartet werden. Für die Evaluation komplexer Interventionen wurden auch schon Empfehlungen gegeben (Campbell et al. 2000; Oakley et al. 2006). Der Ansatz einer Begriffsklärung und ihre Dokumentation, z. B. in der BZgA-Publikation »Leitbegriffe der Gesundheitsförderung und Prävention«, wird darüber hinaus angeregt.

Literatur

Campbell M, Fitzpatrick R, Haines A et al. (2000) Framework for design and evaluation of complex interventions to improve health. BMJ 321: 694–696

Lorenz E (1993) The butterfly effect. In: Lorenz E. The Essence of Chaos. Appendix 1. Seattle, University of Washington Press, S 181–184

Lorenz EN (1963) Deterministic nonperiodic flow. J Atmos Sci 20: 130–141

Oakley A, Strange V, Bonell C et al. and RIPPLE Study Team (2006) Process evaluation in randomised controlled trials of complex interventions. BMJ 332: 413–416

Rittel H, Webber M (1973) Dilemmas in a general theory of planning. Policy Sciences 4: 155–169

West GB (2012) The importance of quantitative systemic thinking in medicine. Lancet 379: 1551–1559

Autorenverzeichnis

Bödeker, Dr. Wolfgang
Bundesverband der Betriebskrankenkassen, Essen
BoedekerW@bkk-bv.de

Braunegger-Kallinger, Mag. Gudrun
Fonds Gesundes Österreich, Wien
Gudrun.Braunegger@goeg.at

Elkeles, Prof. Dr. Thomas
Hochschule Neubrandenburg, Neubrandenburg
Elkeles@hs-nb.de

Hart, Diana
Bundesministerium für Gesundheit, Berlin
Diana.Hart@bmg.bund.de

Kirschner, Dr. Wolf
Forschung, Beratung + Evaluation GmbH, Berlin
Wolf.Kirschner@fb-e.de

Kliche, Dr. Thomas
Hochschule Magdeburg-Stendal, Stendal
Thomas.Kliche@hs-magdeburg.de

Kuhn, Dr. Joseph
Bayerisches Landesamt für Gesundheit und Lebensmittelsicherheit, Oberschleißheim
Joseph.Kuhn@lgl.bayern.de

Kurth, Dr. Bärbel-Maria
Robert Koch-Institut, Berlin
KurthB@rki.de

Lampert, Dr. Thomas
Robert Koch-Institut, Berlin
LampertT@rki.de

Lenz, Dr. Matthias
Universität Hamburg, Hamburg
Matthias.Lenz@uni-hamburg.de

Lorenzen, Dr. Hans-Peter
Deutsche Gesellschaft für Evaluation, Mainz
Hans-Peter.Lorenzen@t-online.de

Maschewsky-Schneider, Prof. Dr. Ulrike
Charité, Berlin School of Public Health, Berlin
Ulrike.Maschewsky-Schneider@charite.de

Meyer, Prof. Dr. Gabriele
Universität Witten/Herdecke, Witten
Gabriele.Meyer@uni-wh.de

Mühlhauser, Prof. Dr. Ingrid
Universität Hamburg, Hamburg
Ingrid_Muehlhauser@uni-hamburg.de

Peinhaupt, MBA, Mag. Christa
Fonds Gesundes Österreich, Wien
Christa.Peinhaupt@goeg.at

Rabe, Nicole
Forschung, Beratung + Evaluation GmbH, Berlin
Nicole.Rabe@fb-e.de

Reisig, Dr. Veronika
Bayerisches Landesamt für Gesundheit und Lebensmittelsicherheit, Oberschleißheim
Veronika.Reisig@lgl.bayern.de

Rohrauer-Näf, MPH, Mag. Gerlinde
Fonds Gesundes Österreich, Wien
Gerlinde.Rohrauer@goeg.at

Schmitt, Dr. Britta
Bundesanstalt für Arbeitsschutz und Arbeitsmedizin, Berlin
Schmitt.Britta@baua.bund.de

Simek, Mag. Monika
Ludwig Boltzmann Institut Health Promotion Research, Wien
Monika.Simek@lbihpr.lbg.ac.at

Thelen, Martina
Robert Koch-Institut, Berlin
ThelenM@rki.de

Trojan, Prof. Dr. Dr. Alf
Institut für Medizinische Soziologie, Sozialmedizin und Gesundheitsökonomie, Universitätsklinikum Hamburg-Eppendorf, Hamburg
Trojan@uke.de

Uhl, Dr. Alfred
Anton-Proksch-Institut, Wien
Alfred.Uhl@api.or.at

Waldherr, Mag. Dr. Karin
Ludwig Boltzmann Institut Health Promotion Research, Wien
Karin.Waldherr@lbihpr.lbg.ac.at

Werner-Schlechter, Barbara
Bundesministerium für Gesundheit, Berlin
Barbara.Werner-Schlechter@bmg.bund.de

Wildner, Prof. Dr. Manfred
Bayerisches Landesamt für Gesundheit und Lebensmittelsicherheit, Oberschleißheim, Pettenkofer School of Public Health
Manfred.Wildner@lgl.bayern.de

Winkler, Dr. Ute
Bundesministerium für Gesundheit, Berlin
Ute.Winkler@bmg.bund.de

Ziese, Dr. Thomas
Robert Koch-Institut, Berlin
ZieseT@rki.de

In the field of prevention there are often complex intervention programmes in which many players are working on different aspects of content and using different methods under a joint organizational roof to achieve a joint, higher goal. Although a tried-and-tested set of methodological instruments is available for evaluating individual preventive measures, to date there are no methodological standards for evaluating complex intervention programmes. Consideration of the methodological demands that would have to be made on the evaluation of complex interventions in prevention is still in its infancy in Germany.

A workshop conducted by the Robert Koch Institute and the Bavarian Health and Food Safety Authority in 2011 discussed the problem of evaluating complex interventions using specific examples. The key questions covered included:

- What characterizes complex interventions and what are the consequences of this for their evaluation?
- In what ways does the evaluation of complex interventions differ from the evaluation of individual measures?
- Is the evaluation of complex interventions more than adding together the evaluations of individual measures?
- Does the evaluation of complex interventions have specific target dimensions and, if so, what are they?
- Can concepts of learning systems make a contribution to the evaluation of complex intervention programmes?
- Are there generalizable aspects of the evaluation of complex interventions that could form a basis for recommendations?

This booklet documents the papers presented at this workshop and additional articles on the evaluation of complex interventions. It is aimed both at scientists working in evaluation research and at practitioners who plan, carry out or evaluate complex intervention programmes.

In der Prävention gibt es häufig komplexe Interventionsprogramme, bei denen zu einer übergeordneten Zielsetzung unter einem gemeinsamen organisatorischen Dach viele Akteure mit verschiedenen Methoden an verschiedenen inhaltlichen Aspekten arbeiten. Während die Evaluation von Einzelmaßnahmen der Prävention auf bewährte methodische Instrumentarien zurückgreifen kann, gibt es zur Evaluation komplexer Interventionsprogramme bisher keine methodischen Standards. Die Reflexion der methodischen Anforderungen an die Evaluation komplexer Interventionen in der Prävention steht in Deutschland noch am Anfang.
Im Rahmen eines Workshops, den das Robert Koch-Institut und das Bayerische Landesamt für Gesundheit und Lebensmittelsicherheit 2011 durchgeführt haben, wurde die Problematik der Evaluation komplexer Interventionen anhand konkreter Beispiele diskutiert. Leitfragen dabei waren:

- Was charakterisiert komplexe Interventionen und was folgt daraus für ihre Evaluation?
- Was unterscheidet die Evaluation komplexer Interventionen von der Evaluation einzelner Maßnahmen?
- Ist die Evaluation komplexer Interventionen mehr als die Addition der Evaluation einzelner Maßnahmen?
- Hat die Evaluation komplexer Interventionen besondere Zieldimensionen und wenn ja, welche?
- Können Konzepte lernender Systeme etwas zur Evaluation komplexer Interventionsprogramme beitragen?
- Gibt es verallgemeinerbare Aspekte in der Evaluation komplexer Interventionen, die Grundlage von Empfehlungen sein könnten?

Der vorliegende Band dokumentiert die Vorträge dieses Workshops und weitere Beiträge zum Thema Evaluation komplexer Interventionen. Er wendet sich sowohl an Wissenschaftler und Wissenschaftlerinnen, die in der Evaluationsforschung tätig sind, als auch an Praktiker und Praktikerinnen, die komplexe Interventionsprogramme planen, durchführen oder evaluieren.

ISBN 978-3-89606-215-4

Das Robert Koch-Institut ist ein Bundesinstitut
im Geschäftsbereich des Bundesministeriums für Gesundheit