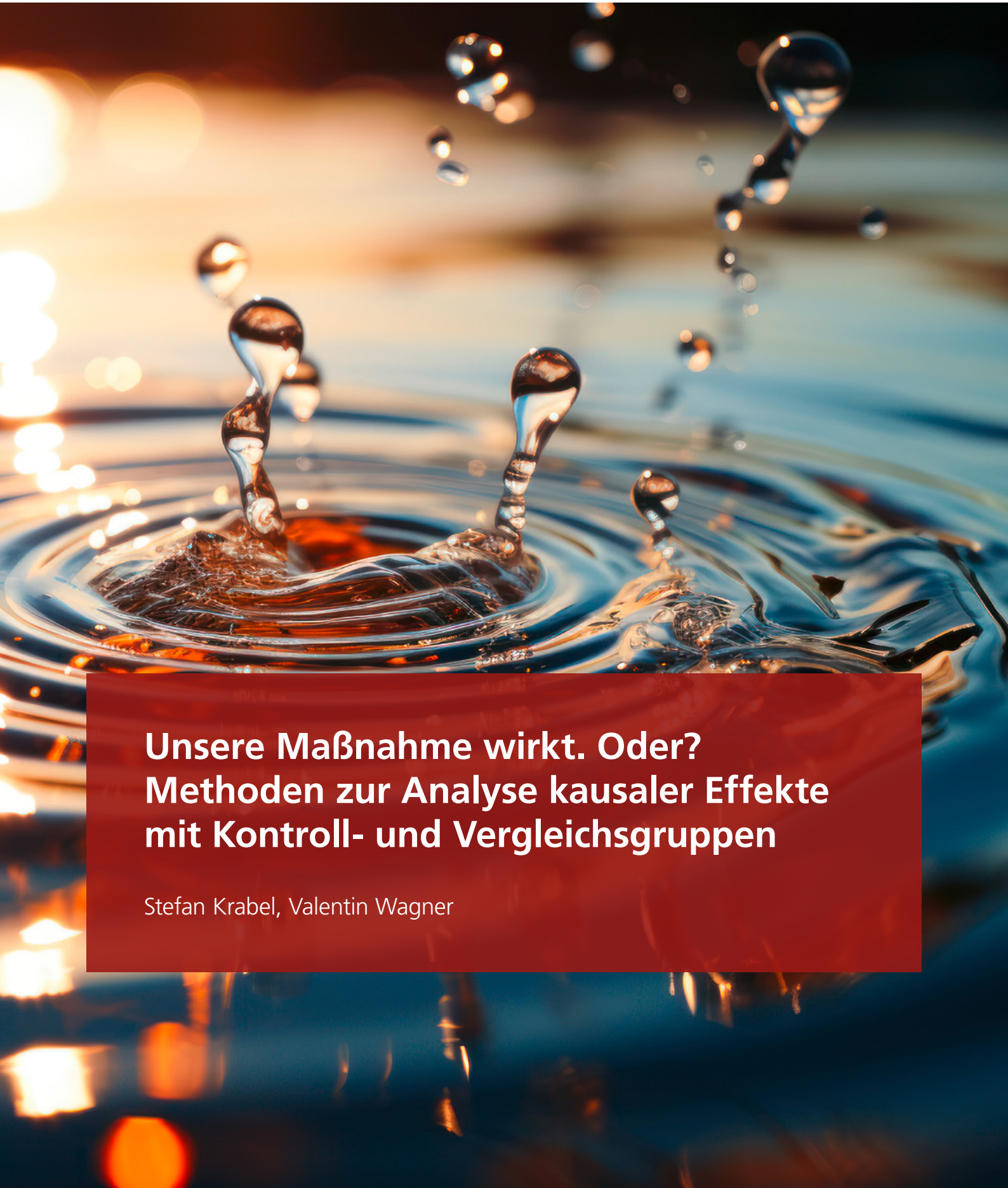
INSTITUT FÜR
INNOVATION UND
TECHNIK



Unsere Maßnahme wirkt. Oder? Methoden zur Analyse kausaler Effekte mit Kontroll- und Vergleichsgruppen

Stefan Krabel, Valentin Wagner

Impressum

Herausgeber
Prof. Dr. Wittpahl
Institut für Innovation und Technik (iit)
in der VDI/VDE Innovation + Technik GmbH
Steinplatz 1
10623 Berlin
Tel.: +49 30 310078 5507
Fax: +49 30 310078 216
E-Mail: info@iit-berlin.de
www.iit-berlin.de

Autoren
Dr. Stefan Krabel
Dr. Valentin Wagner

Layout
Poli Quintana

Bildrechte
Sebastian – stock.adobe.com

DOI: 10.23776/2024_11
Berlin, Juni 2024

Zitation
Krabel, Stefan; Wagner, Valentin (2024): Unsere Maßnahme wirkt. Oder? Methoden zur Analyse kausaler Effekte mit Kontroll- und Vergleichsgruppen. Institut für Innovation und Technik (iit).

Inhalt

1 Einleitende Worte – und die Frage nach der Wirksamkeit von Fördermaßnahmen

Im Rahmen der Forschungs- und Innovationsförderung werden öffentliche Mittel zur Durchführung von Forschungsprogrammen umgesetzt, um bestimmte (politische und gesellschaftliche) Ziele zu erreichen – etwa die stärkere Nutzung von Elektrofahrzeugen, der vermehrte Einsatz digitaler Technologien in der Bildung oder die Vernetzung unterschiedlicher Fachdisziplinen. In diesem Zusammenhang hat die Expertenkommission Forschung und Innovation (EFI) kürzlich darauf hingewiesen, dass in vielen Fällen die Wirkung der in Evaluationen untersuchten Maßnahmen methodisch nicht verlässlich gemessen werden (EFI 2024, S. 40):

„Viele der im Auftrag der Bundesregierung durchgeführten Evaluationsstudien zu Maßnahmen der Forschungs- und Innovationspolitik (F&I-Politik) lassen keine Rückschlüsse darauf zu, ob die beobachteten Entwicklungen tatsächlich auf die untersuchten Politikmaßnahmen zurückzuführen sind. Wesentlicher Grund hierfür ist, dass Evaluationsstudien häufig nicht den methodischen Anforderungen an eine Kausalanalyse genügen, nicht zuletzt, weil die Voraussetzungen für den sachgerechten Einsatz geeigneter Methoden nicht in jedem Fall erfüllt sind. Das fehlende Wissen über die Wirkung von Maßnahmen erschwert ein systematisches und evidenzbasiertes Politiklernen".

Damit greift Die EFI-Kommission eine Forderung der Europäischen Kommission auf, die Evaluationen mit kontrafaktischer Evidenz und Kontrollgruppenansätzen seit Jahren stärker fordert (z. B. im Rahmen des European Structural Funds; siehe Europäische Kommission 2023) und auch Leitfäden zur Umsetzung derartiger Methoden in Evaluationen verfasst hat (siehe z. B. Europäische Kommission 2022).

Doch wie können Evaluatorinnen und Evaluatoren die Wirkung einer Maßnahme methodisch hinreichend verlässlich messen? Um evidenzbasierte Aussagen zur kausalen Wirkung von Maßnahmen treffen zu können, ist eine zentrale wissenschaftliche Methode, Vergleiche zwischen geeigneten Gruppen durchzuführen. Bei diesem Ansatz wird zwischen sogenannten Treatment- und Kontrollgruppen unterschieden. Die erstgenannte Gruppe bekommt eine spezifische Behandlung („Treatment"), z. B. ein Medikament oder eine staatliche Förderung, die Kontrollgruppe bekommt diese nicht. Mit Hilfe geeigneter statistischer Verfahren werden dabei Gruppenvergleiche vorgenommen, die eine kausale Analyse von Effekten – also Wirkung – der Maßnahme ermöglichen.

Die Methodik des Vergleichs- und Kontrollgruppenansatzes kann in unterschiedlichen Bereichen genutzt werden. Sie stammt aus der quantitativen Forschung und ist in den Naturwissenschaften, insbesondere der Medizin, etabliert. Sie kommt aber auch in den Geisteswissenschaften zum Einsatz und kann u.a. für die Evaluation politischer Maßnahmen genutzt werden (siehe Wangler 2015).

Ziel des vorliegenden Beitrags ist es, Methoden zur Bildung von Kontroll- und Vergleichsgruppen sowie Methoden zur Analyse kausaler Effekte darzustellen. Dabei sollen die Anforderungen an Kontroll- und Vergleichsgruppenanalysen herausgearbeitet werden, denen genügt werden muss, um mit diesen Methoden kausale Effekte „verzerrungsfrei" – also ohne systematischen Fehler – messen zu können. Ferner wird auf die Rahmenbedingungen bei der Datenverfügbarkeit und dem Datenzugang eingegangen, da auch diese Aspekte häufig eine Herausforderung bei der Evaluation von innovationspolitischen Maßnahmen darstellen. Abschließend werden Möglichkeiten skizziert, wie Fördermaßnahmen aufgesetzt werden könnten, um im Anschluss an die Förderung eine kausale Wirkung der Maßnahme – mit Kontroll- und Vergleichsgruppen – zu ermöglichen.

2 Identifizierung von Kontroll- und Vergleichsgruppen

2.1 Der „Gold-Standard“: Randomisierte kontrollierte Studie

Randomisierte kontrollierte Studien (Randomized Controlled Trials = RCTs) gelten in der empirisch-quantitativen Wissenschaft als Gold-Standard für die eindeutige Identifizierung von kausalen Zusammenhängen.[1] Zentraler und zugleich wichtigster Bestandteil von RCTs ist die *zufällige* Zuteilung von Beobachtungseinheiten (z. B. Individuen oder Institutionen) in eine oder mehrere Treatment- und eine Kontrollgruppe. Durch die zufällige Zuteilung soll gewährleistet werden, dass beide Gruppen im statistischen Sinne identisch sind, d.h. individuelle Unterschiede der Beobachtungseinheiten nivellieren sich durch ihre beliebige Zusammenstellung.[2] Anschließend erhält die Treatmentgruppe eine bestimmte Behandlung (dies kann u.a. eine bestimmte Ausstattung oder eine bestimmtes Medikament sein), die Kontrollgruppe erhält diese dagegen nicht.

Der Effekt der Behandlung (unabhängige Variable) auf das Merkmal von Interesse (abhängige Variable) kann durch einen Vergleich der Treatmentgruppe mit der Kontrollgruppe gemessen werden (siehe z. B. Arni 2012). Dabei sollte die RCT idealerweise so gestaltet werden, dass kein Informationsaustausch und keine Interaktionen zwischen den beiden Gruppen stattfinden, um die Möglichkeit von Spillover-Effekten (Übertragungseffekte) zu vermeiden. Von einem Spillover-Effekt ist die Rede, wenn das geänderte Verhalten (aufgrund der Maßnahme) in der Treatmentgruppe durch Interaktionen einen Effekt auf die Kontrollgruppe hat. Zudem werden durch das Auslassen von Informationsflüssen der sogenannte Experimenter-Demand-Effect und der sogenannte Hawthorne-Effekt vermieden. Diese beiden Effekte bezeichnen ein verzerrtes Verhalten der Individuen, dass aus ihrem Wissen resultiert, an einem Experiment teilzunehmen oder sich in der Treatmentgruppe zu befinden. In diesem Fall können die Individuen beispielsweise überlegen, welches (ggf. sozial erwünschte) Ergebnis der Experimentator beobachten möchte und ändern ihr Verhalten dementsprechend, wobei diese Verhaltensänderung nicht notwendigerweise mit Absicht erfolgt.

Bei der Randomisierung bzw. der zufälligen Einteilung in Kontroll- und Treatmentgruppe sollten je nach Stichprobengröße und Rahmenbedingungen verschiedene Ansätze gewählt werden. Ein Beispiel für ein Randomisierungsverfahren ist z. B. „pure randomization“, d.h. die Probanden werden per Lotterie in zwei oder mehr Gruppen aufgeteilt, indem beispielsweise Münzwürfe oder Zufallsnummerngeneratoren verwendet werden. Dieses Verfahren bietet sich insbesondere an, wenn die Stichprobe, die auf die beiden Gruppen aufgeteilt werden soll, relativ groß ist. Eine andere Methode ist die sogenannte „stratifizierte Randomisierung“. Dabei werden die Personen nach bestimmten Charakteristika (z. B. Geschlecht, Alter, Herkunftsregion) in Schichten eingeteilt. Anschließend wird aus jeder Schicht per Zufall ausgelost, wer am Experiment teilnimmt und in welche Treatmentgruppe er/sie kommt.[3] Die stratifizierte Randomisierung hilft mögliche Störfaktoren bei der Interpretation der Studienergebnisse zu minimieren.

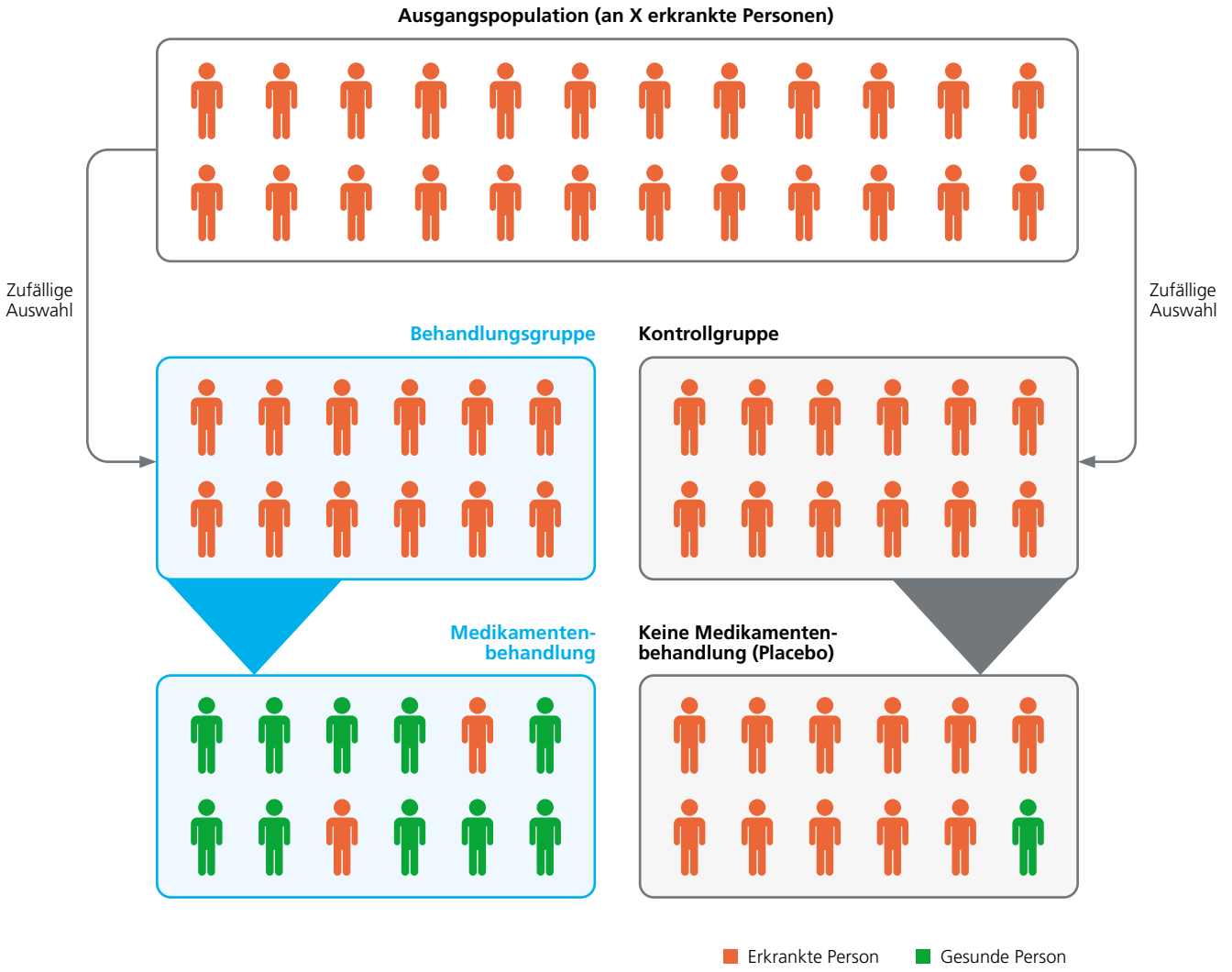
Ein einfaches Beispiel für eine randomisierte kontrollierte Politikintervention ist die Einführung von Lehrassistenten in Dänemark (Andersen et al. 2020). Die Sozialdemokraten versprachen in den Parlamentswahlen im Herbst 2011 die Einführung von Lehrassistenten in allen Grundschulklassen, wenn sie die Wahl gewinnen. Nach der Wahl wurde jedoch nicht sofort in allen Grundschulklassen eine Lehrassistenz eingeführt, sondern ein Experiment initiiert, um die Auswirkungen der Lehrassistenten kausal evaluieren zu können. Es wurden zufällig 105 Schulen im ganzen Land ausgewählt in deren sechsten Klassen Lehrassistenten eingeführt wurden (Treatmentgruppe) und die schulischen Leistungen dieser Kinder mit Sechstklässlern in Schulen ohne Lehrassistenten verglichen (Kontrollgruppe). Die Studienergebnisse zeigten einen positiven und nachhaltigen Effekt von Lehrassistenten auf die schulische Leistung auf. Wären die Lehrassistenten gleichzeitig in allen Grundschulen eingeführt worden, wäre es nicht möglich gewesen den Effekt der Lehrassistenten von anderen zeitgleichen Effekten oder Politikmaßnahmen zu trennen.

Neben der randomisierten kontrollierten Studie geben auch *natürliche Experimente* Aufschluss über kausale Zusammenhänge. Natürliche Experimente sind Fälle bzw. Beispiele, in denen die „Natur“ eine zufällige Aufteilung in eine Treatment- und eine Kontrollgruppe vornimmt, d.h. die zufällige Einteilung in beide Gruppen lag nicht in den Händen der Forschenden, sondern die Randomisierung erfolgte zufällig durch Kontextfaktoren in der Realität.

Ein historisches Beispiel ist die Untersuchung eines Ausbruchs einer Cholera-Epidemie des britischen Chirurgen John Snow aus

1 2019 wurde der Alfred-Nobel-Gedächtnispreis für Wissenschaften u.a. für die Verwendung von RCTs in der Entwicklungsökonomie vergeben.
2 Dies ist der Fall, wenn die Anzahl an Beobachtungseinheiten groß genug ist. Bei einer kleineren Anzahl an Beobachtungseinheiten müssen ggf. spezielle Randomisierungsverfahren angewendet werden.
3 Weitere Randomisierungsverfahren sind beispielsweise das pairwise matching, block.randomization oder re-randomization. Diese Verfahren werden in diesem Papier jedoch nicht näher behandelt.

Auswahl und Untersuchung des Behandlungserfolges mit Behandlungs- und Kontrollgruppe



dem Jahr 1855: es konnte gezeigt werden, dass sich das Gebiet in London, in dem die Cholera-Epidemie grassierte, mit dem Versorgungsgebiet einer (zufällig) verunreinigten Wasserpumpe deckte. Die Versorgungsgebiete anderer Wasserquellen waren bei Weitem nicht so stark betroffen (Snow 1855). Aus dieser Deckung des zufälligen (natürlichen) Unterschieds in der Wasserqualität mit der regionalen Verbreitung (regionale Infektionsrate bzw. Todesrate nach Scharlach und Fieber) der Epidemie konnte abgeleitet werden, dass die Krankheit auf im Wasser befindliche Mikroorganismen zurückzuführen war.

2.2 Quasi-experimentelle Ansätze zur Erfassung von Kontrollgruppen

Die sogenannte *quasi-experimentelle Studie* oder *nicht-randomisierte kontrollierte Studie* greift zur kausalen Prüfung einer Hypothese auf Gruppen zurück, die nicht zufällig gebildet wurden bzw. werden konnten (keine Randomisierung), sondern nachträglich aus einer Gesamtpopulation ermittelt wurden. Diese Bildung kann mit sogenannten Matching-Verfahren vorgenommen werden. Ein Beispiel dafür ist das Propensity Score Matching (Austin 2014), mit dem aus großen Datensätzen sta-

tistische „Zwillingspaare" gebildet werden, die sich idealerweise hinsichtlich eines interessierenden Merkmals (analog zum Treatment) unterscheiden, ansonsten aber ähnliche Charakteristika aufweisen (siehe Baser 2006; Austin 2014).[4]

Alternativ zum Matching können bestimmte Schwellenwerte genutzt werden, um möglichst gut vergleichbare Gruppen zu erzeugen (sogenannter „Regression-Discontinuity-Ansatz"). Beispielsweise kann es eine Leistungsmetrik geben, anhand derer ein Stipendium an 100 Personen vergeben wird. Eine Möglichkeit für eine Aufteilung in Gruppen wäre hierbei, Personen, die gerade noch die geforderten Leistungen erfüllen – und entsprechend eine Förderung erhalten haben – zu vergleichen mit Personen, die sehr knapp die Leistung nicht erfüllen – und entsprechend keine Förderung erhalten. In diesem Fall wären beide Gruppen ähnlich leistungsstark, so dass ein Vergleich der Gruppen zur Analyse der Wirkung des Stipendiums genutzt werden kann. Die Annahme dabei ist, dass die Unterschiede der Personen in ihren Fähigkeiten sehr gering sind und eine spätere Auswirkung des Stipendiums (z. B. auf späteres Einkommen) auf diese Förderung zurückzuführen ist. Somit wird eine Diskontinuität oder Unstetigkeit in einer beobachteten Kontrollvariable genutzt, um eine „fast zufällige" Zuteilung in die Treatment- oder Kontrollgruppe vorzunehmen (Imbens und Lemieux 2008). Ein Überblick über Methoden zur Bildung dieser Gruppen findet sich in Tabelle 1.

Überblick über Methoden zur Bildung von Treatment- und Kontrollgruppe

Methode	Beschreibung	Datenbedarf
Randomisierte kontrollierte Studie	Eine Gesamtpopulation wird zufällig in zwei Gruppen geteilt: Die Treatmentgruppe bekommt ein bestimmtes „Treatment" bzw. eine bestimmte Behandlung (dies kann u.a. eine bestimmte Ausstattung oder ein bestimmtes Medikament sein), die Kontrollgruppe erhält diese dagegen nicht. Die zufällige Verteilung kann beispielsweise mit Hilfe von Zufallsgeneratoren, Losentscheid oder Münzwurf erfolgen.	Schon vor der Datenerhebung wird ein kontrolliertes Experiment geplant, in dem der Effekt einer bestimmten Variable auf eine Zielgröße gemessen werden soll. Es besteht der Bedarf, Datenerhebungen und eine Zuteilung in Gruppen durchführen zu können – die Forschungsfrage darf also nicht im Nachgang zur Datenerhebung formuliert werden.
Analyse natürlicher Experimente	Eine reale Beobachtung ist vergleichbar mit einer experimentellen Randomisierung; es gibt zwei Gruppen, die sich nur in einer interessierenden Variable unterscheiden. Die Zuteilung wurde von der „Natur" und nicht den Forschenden vorgenommen.	Zu den zwei Gruppen eines natürlichen Experiments liegen alle interessierenden Informationen vor.
Quasi-experimentelle Studie	Mit Hilfe von statistischen Verfahren können aus großen bestehenden Datensätzen zwei Gruppen gebildet werden, die sich in einer Variable unterscheiden und ansonsten sehr ähnlich sind. Ein häufiges Verfahren hierzu ist das sogenannte „Propensity Score Matching". Bei diesem Verfahren werden Paare von möglichst identischen Personen – hinsichtlich ausgewählter Variablen des Datensatzes – gebildet, die miteinander verglichen werden können. Es erfolgt keine zufällige Einteilung in Kontroll- und Treatmentgruppe.	Benötigt werden Datensätze mit großer Fallzahl und hinreichend vielen interessierenden Variablen, die ein Matching nach wichtigen Kriterien ermöglichen. Unterschiede zwischen den Gruppen können nicht auf nicht-beobachtbare Variablen zurückgeführt werden.

Tabelle 1: Überblick über Methoden zur Bildung von Treatment- und Kontrollgruppe (Quelle: Institut für Innovation und Technik)

4 Für das Erstellen der Zwillingspaare gibt es eine Vielzahl von Verfahren, beispielsweise (1) Nächste-Nachbarn, (2) Caliper und Radius, (3) Stratifizierung und (4) kernbasierte Verfahren.

3 Statistische Verfahren in der Vergleichsgruppenanalyse

3.1 Einfache statistische Tests

Einen ersten Aufschluss über die Wirkung einer Maßnahme ergeben einfache statistische Tests, die Gruppenvergleiche vornehmen und eine sogenannte Signifikanzanalyse vornehmen. Dabei wird untersucht, ob beobachtete Unterschiede in der interessierenden Variable systematisch und nicht zufällig auftreten. Ein Beispiel: Forschende haben die Größe von 100 Pflanzen jeder Art gesammelt und wollen nun die unterschiedliche Größe von zwei ausgewählten Pflanzengattungen analysieren und schauen, ob der Größenunterschied der beiden Gattungen statistisch signifikant ist. Sie kommen zu dem Ergebnis, dass die durchschnittliche Größe der ersten Pflanzenart 10 Zentimeter beträgt, während die durchschnittliche Größe der zweiten Pflanzenart 15 Zentimeter beträgt. Um die Signifikanz dieses Unterschiedes zu ermitteln, d.h. zu untersuchen, ob der Unterschied von 5 Zentimetern systematisch oder rein zufällig ist, kann beispielsweise zunächst die Teststatistik „t" ermittelt werden: Hierbei handelt es sich um ein statistisches Maß, das basierend auf den Mittelwerten der Gruppen, der Stichprobengröße und der Varianz der Daten errechnet wird. Der t-Wert wird dann mit einem kritischen Wert aus der t-Verteilung verglichen, um festzustellen, ob der beobachtete Unterschied zwischen den Gruppen signifikant ist oder nur auf zufälligen Variationen beruht. Je höher der t-Wert ist, desto höher ist das Signifikanz-Niveau, d.h. die Wahrscheinlichkeit einer fehlerhaften Messung ist geringer.

Die Wahl des Verfahrens ist dabei abhängig von der Ausprägung der betrachteten Variable (Zahlenwert, Rangfolge, qualitative Daten). Einige der häufigsten Methoden sind im Folgenden kurz skizziert (für detailliertere Auswertungen siehe z. B. Kuß et. al. 2014):

- t-Test: Der t-Test wird verwendet, um festzustellen, ob der Mittelwert der Zielgröße in der Kontrollgruppe signifikant von dem Mittelwert der Zielgröße in der Treatmentgruppe (generell: zwischen unabhängigen Stichproben) abweicht. Hierbei wird davon ausgegangen, dass die Daten normalverteilt sind und die Varianzen zwischen den Gruppen gleich groß sind.

- Mann-Whitney-U-Test: Der Mann-Whitney-U-Test wird verwendet, um festzustellen, ob die Ränge einer Variablen in einer Gruppe signifikant von den Rängen in einer anderen Gruppe abweichen. Dieser Test wird angewendet, wenn die Daten nicht normalverteilt sind.

- Chi-Quadrat-Test: Der Chi-Quadrat-Test wird verwendet, um festzustellen, ob es einen statistisch signifikanten Zusammenhang zwischen zwei qualitativen (bzw. kategorialen) Variablen gibt.

- ANOVA: Die ANOVA (Analyse der Varianz) wird verwendet, um festzustellen, ob es signifikante Unterschiede in den Mittelwerten einer abhängigen Variablen zwischen drei oder mehr Gruppen gibt. Der Test vergleicht die Varianz zwischen den Gruppen mit der Varianz innerhalb der Gruppen.

3.2 Multivariate Analysen

Multivariate Verfahren werden in der Statistik angewendet, wenn der Zusammenhang mehrerer Variablen untersucht wird, etwa um komplexe Zusammenhänge und Muster zu erkennen. Sie kommen zum Einsatz, wenn die abhängige Variable von mehreren unabhängigen Variablen beeinflusst wird – oder, wenn mehrere Variablen miteinander verglichen werden sollen. Multivariate Verfahren werden häufig in Bereichen wie der Marktforschung, der medizinischen Forschung, der Psychologie, der Ökonometrie und anderen Bereichen angewendet, um komplexe Daten zu analysieren und Schlussfolgerungen zu ziehen.

Ein bekanntes Beispiel für die multivariate Analyse von einer Kontroll- und Treatmentgruppe in einem natürlichen Experiment ist die Untersuchung des Effekts der Einführung eines Mindestlohns auf die Arbeitslosigkeit von Card und Krueger (1993). Im April 1992 wurde der Mindestlohn in New Jersey (= Treatmentgruppe) von 4,25 Dollar auf 5,05 Dollar angehoben, der Mindestlohn im Nachbarstaat Pennsylvania (= Kontrollgruppe) blieb hingegen konstant bei 4,25 Dollar. Die Autoren führten Datenerhebungen in Fast Food-Restaurants im Februar und November des Jahres 1992 (also vor und nach der Änderung des Mindestlohns) in beiden Staaten durch, um zu untersuchen, ob die Einführung des Mindestlohns einen Effekt auf die Arbeitslosenzahlen in New Jersey hat. Mit Hilfe von multivariaten Regressionsanalysen wurden anschließend Gruppenvergleiche herausgearbeitet: Dabei wurde als abhängige Variable (=Zielgröße) die Veränderung der Anzahl an Arbeitsplätzen in Fast Food-Restaurants im jeweiligen Staat betrachteten, als erklärende Variablen wurden unterschiedliche Merkmale, wie etwa die Anzahl der Vollzeitäquivalente der Beschäftigten in dem Restaurant oder andere Charakteristika der unterschiedlichen Fast-

Food-Ketten, berücksichtigt. Entscheidend bei der Analyse ist, dass zusätzlich eine binäre Variable (=Behandlungsvariable) aufgenommen wurde, die den Wert von „1" annahm, wenn das Fast Food-Restaurant einen Mindestlohn zahlen musste und einen Wert von „0" annahm, wenn kein Mindestlohn gezahlt wurde. Die Untersuchung fand keine signifikante Wirkung des Mindestlohns auf die Anzahl an Arbeitsplätzen.

Neben der multivariaten Regressionsanalyse kann unter gewissen Bedingungen auch die sogenannte „Difference-in-Differences-Schätzmethode" eingesetzt werden. Diese ist insbesondere geeignet, wenn für die Kontroll- und Treatmentgruppe Daten im Rahmen einer Längsschnittstudie (wiederkehrende Erhebungen zu regelmäßigen Zeitpunkten) vorliegen. Diese Methode wird häufig verwendet, um den kausalen Effekt einer Intervention oder einer Behandlung auf eine Gruppe im Vergleich zu einer anderen Gruppe zu messen. Dabei werden die absolute oder die relative Veränderung der abhängigen Variable vor und nach der Intervention in beiden Gruppen berechnet und miteinander verglichen.

Der große Vorteil dieser Methode ist, dass sich Kontroll- und Treatmentgruppe in (mehreren) Merkmalen zum Zeitpunkt vor der Intervention unterscheiden dürfen. Als erster Schritt wird in dieser Methode also ermittelt, wie groß der Unterschied zwischen Kontroll- und Treatmentgruppe in der Ausgangslage (vor der Intervention) ist. Nachdem die Intervention erfolgt ist, wird zu einem späteren Zeitpunkt erneut der Unterschied zwischen der Kontroll- und der Treatmentgruppe gemessen. Anschließend werden diese, zu unterschiedlichen Zeitpunkten, ermittelten Unterschiede miteinander verglichen und berechnet, inwiefern sie voneinander abweichen. Ist die Abweichung signifikant, wird dies der kausale Effekt der Intervention.

Ferner gibt es in der Statistik und Ökonometrie zahlreiche weitere Methoden und Verfahren, die zur Identifizierung kausaler Effekte und „Messung verzerrungsfreier" Effektstärken beitragen, wenn ein kontrolliertes Experiment nicht möglich ist, oder das Experiment nicht korrekt durchgeführt werden konnte. Wenn beispielsweise untersucht werden soll, welchen Effekt Bildung auf das Einkommen hat, ist dies nicht problemlos möglich, da andere Faktoren wie beispielsweise Fleiß oder sozialer Status dazu führen können, dass eine Person sowohl viel Zeit in Bildung investiert als höhere Einkommen bezieht. Damit würde eine Analyse, in der Bildung (Variable X) in Beziehung zum Einkommen (Variable Y) gesetzt wird (statistisch ausgedrückt: X wird auf Y regressiert), zu falschen bzw. ungenauen Schlussfolgerungen führen. In diesen Fällen kann das Problem mit Hilfe einer instrumentalen Variablenschätzung gelöst werden. Der Grundgedanke der instrumentalen Variablenschätzung besteht darin, eine (zusätzliche) Variable Z (z. B. Nähe des Wohnortes zu einer Bildungseinrichtung) zu identifizieren, die mit der unabhängigen Variable X (Bildungsniveau) korreliert, aber nicht mit der abhängigen Variable Y (Einkommen). In anderen Worten: Variable Z hat einen Effekt auf Y, aber nur durch seinen Effekt auf X. Zudem ist anzunehmen, dass die Nähe zu einer Bildungseinrichtung keinen Effekt auf Fleiß oder sozialen Status hat. Diese zusätzliche Variable wird als *Instrument* bezeichnet und dient dazu, die unerwünschte Störung durch gleichzeitig wirkende Faktoren zu eliminieren. Zur Schätzung der Kausalität zwischen X und Y zu dient dabei eine Regressionsanalyse, bei der Y auf Z und Z auf X regressiert wird. Die Parameter der geschätzten Gleichung geben dann den kausalen Effekt von X auf Y an, wobei die Störung durch andere Faktoren kontrolliert wird (siehe Kleibergen und Zivot 2003; Wangler 2014).

Übersicht über Methoden zu Gruppenvergleichen

Methode	Beschreibung	Anmerkung
Propensity Score Matching	Der Propensity Score (PS) wird als die Wahrscheinlichkeit definiert, mit der eine Analyseeinheit (z. B. eine Person oder ein Unternehmen) eine zu prüfende Behandlung (z. B. Therapie) erhält. Der PS wird in einem ersten Schritt aus den vorhandenen Daten geschätzt, beispielsweise über eine Regression. Im zweiten Schritt erfolgt die Schätzung des eigentlich interessierenden Behandlungseffekts unter Zuhilfenahme des PS.	Beim Matching nach dem PS stehen verschiedene Methoden zur Verfügung, wie z. B. Nächste-Nachbarn-Verfahren, Stratifizierung oder kernbasierten Verfahren.
Einfache statistische Tests zur Analyse von Gruppenvergleichen (Auswahl)		
t-Test	Der t-Test wird verwendet, um festzustellen, ob der Mittelwert einer Variablen in einer Gruppe signifikant von dem Mittelwert einer anderen Gruppe abweicht.	Die Auswahl des Tests erfolgt anhand der Beschaffenheit der in den Fokus genommenen Variable (kontinuierliche Zahl, qualitative Variable, Rangfolge) und der konkreten Fragestellung.
Mann-Whitney-U-Test	Der Mann-Whitney-U-Test wird verwendet, um festzustellen, ob die Ränge einer Variablen in einer Gruppe signifikant von den Rängen in einer anderen Gruppe abweichen.	
Chi-Quadrat-Test	Der Chi-Quadrat-Test wird verwendet, um festzustellen, ob es einen statistisch signifikanten Zusammenhang zwischen zwei qualitativen (bzw. kategorialen) Variablen gibt.	
ANOVA	Die ANOVA (Analyse der Varianz) wird verwendet, um festzustellen, ob es signifikante Unterschiede in den Mittelwerten einer abhängigen Variablen zwischen drei oder mehr Gruppen gibt.	
Multivariate Verfahren zur Analyse von Gruppenvergleichen (Auswahl)		
Difference-in-Differences-Schätzer	Der absolute oder relative Unterschied zwischen beiden Gruppen in der abhängigen Variable zum Zeitpunkt vor der Intervention wird mit dem Unterschied in der abhängigen Variable zum Zeitpunkt nach der Intervention verglichen. Dieser Unterschied (Ausprägung) zwischen den Unterschieden (zeitlich) wird als kausaler Effekt der Intervention interpretiert.	Das Verfahren kann für lineare und logistische Funktionen gleichermaßen durchgeführt werden.
Regressionen mit Instrumentalvariablen	Instrumentalvariablen haben einen Einfluss auf die unabhängige Variable, sind aber nicht direkt mit der abhängigen Variable korreliert. Instrumentalvariablen können in multivariaten Modellen verwendet werden, um Verzerrungen durch endogene Variablen zu reduzieren.	Das Verfahren kann nur angewendet werden, wenn eine Instrumentalvariable identifiziert werden kann, die diesen Anforderungen entspricht.

Tabelle 2: Übersicht über Methoden zu Gruppenvergleichen (Quelle: Institut für Innovation und Technik)

4 Anforderung an Daten und Datenzugang

4.1 Anforderung an Datensätze

In der Praxis ist es bei Evaluationen häufig eine Herausforderung, einen passenden Datensatz für die Analyse zusammenzustellen. Zum einen erfolgt bei der Datenerhebung selten eine Einteilung in Kontroll- und Treatmentgruppen, zum anderen ist auch der Datenzugang begrenzt, um beispielsweise ein Matching und eine damit einhergehende Gruppenzuteilung vorzunehmen (s. Abschnitt 2.2). Um aus bestehenden oder künftig zu erhebenden Datensätzen Treatment- und Kontrollgruppen bilden zu können, sollten die Datensätze die folgenden Voraussetzungen erfüllen:

- Repräsentativität: Die Datensätze stellen eine repräsentative Stichprobe der Gesamtpopulation dar, sodass Ergebnisse „generalisiert" werden können. Dies bedeutet, dass die Merkmale der Teilnehmenden in den Datensätzen die Merkmale der Gesamtpopulation widerspiegeln.

- Hohe Fallzahl / Größe: Die Fallzahl bzw. Größe bedingt, welche Effektstärken als statistisch signifikant identifiziert werden können. Fällt der Effekt einer Intervention beispielsweise eher gering aus, braucht es eine größere Fallzahl an Beobachtungen, um diese kleiner Effektgröße im statistische Sinne auch zu „entdecken".[5]

- Konsistenz: Die Datensätze sollten möglichst konsistent und genau sein. Es ist wichtig sicherzustellen, dass es keine fehlenden oder inkonsistenten Datenpunkte gibt, die die Vergleichbarkeit der Gruppen beeinträchtigen könnten.

- Homogenität: Die Datensätze der Kontrollgruppe und der Experimentalgruppe sind komparativ auswertbar. Dies bedeutet: relevante Faktoren werden in beiden Gruppen beobachtet und sind vergleichbar, die Daten wurden auf dieselbe (oder sehr ähnliche) Weise erhoben.

Erfüllt ein Datensatz diese Kriterien, lassen sich valide Gruppenvergleiche mit Hilfe von einfachen statistischen Tests (siehe Abschnitt 3.1) durchführen. Sollen zusätzlich spezielle multivariate Verfahren angewendet werden, sind ggf. weitere Anforderungen zu erfüllen, die mit der Anwendung der jeweiligen Verfahren verbunden sind.

4.2 Datenzugang

Werden Kontroll- und Vergleichsgruppenansätze nicht schon beim Design von Maßnahmen mitgedacht, kann es schwierig bis unmöglich sein, überhaupt einen geeigneten Datensatz für die Gruppenvergleiche zu identifizieren. Und selbst wenn es passende Datensätze gibt – entweder direkt geeignet für Kontrollgruppenanalysen oder für die Zusammenstellung von geeigneten Kontroll- und Vergleichsgruppen – stehen diese Daten nicht immer für eine Evaluationen zu Verfügung. Das Problem dabei ist juristischer Natur: Viele Evaluationen und Studien, in denen Kontrollgruppenansätze verwendet werden (könnten), werden im Rahmen von Aufträgen vergeben – z. B. von Ministerien oder großen Firmen. Im Rahmen der Auftragsforschung ist der Schutz personenbezogener Daten zu gewährleisten. Im Bundesstatistikgesetz ist dazu geregelt, unter welchen Bedingungen Daten für wissenschaftliche Projekte zur Verfügung gestellt werden dürfen. Konkret ist in §16, Absatz 6 des Bundesstatistikgesetzes (siehe BMJ 2024: §16) u.a. angeführt:

„Für die Durchführung wissenschaftlicher Vorhaben dürfen das Statistische Bundesamt und die statistischen Ämter der Länder Hochschulen oder sonstigen Einrichtungen mit der Aufgabe unabhängiger wissenschaftlicher Forschung (1) Einzelangaben übermitteln […]".

Dies bedeutet, dass Mikrodaten – die oft notwendig sind, um Repräsentativität herzustellen und geeignete Gruppen bilden zu können – zwar für wissenschaftliche Arbeiten an Hochschulen übermittelt werden dürfen, nicht aber für Auftragsforschung – also nicht an Unternehmen der gewerblichen Wirtschaft.

Aus Sicht der evidenzbasierten Politikberatung wäre es generell wünschenswert, wenn der Datenzugang – auch von amtlichen Daten – für die Auftragsforschung (in anonymisierter Form) erlaubt wird und nicht zweck- und adressatengebunden ist. Insbesondere wenn Daten mit Steuergeldern erhoben wurden, sollten diese Daten auch öffentlich zur Verfügung stehen und für Evaluationen genutzt werden können. So können u.a. auch ressourcenaufwendige Doppelerhebungen vermieden werden und dem Grundsatz der Datensparsamkeit gerecht werden.

5 Es gibt statistische Verfahren (sogenannte Power Calculations) mit denen – ex post oder ex ante einer Intervention – berechnet werden kann, welche Fallzahl bzw. Größe benötigt wird/wurde, um Effektgrößen erkennen zu können.

5 Kontrollgruppenanalysen in Evaluationen und Randomisierung in Förderformaten

5.1 Anwendungsmöglichkeiten von Kontrollgruppenanalysen in Evaluationen

Die Anwendung von Kontrollgruppenvergleichen in Evaluationen ist abhängig von der jeweiligen Datenlage und -verfügbarkeit. Der „theoretische Idealfall" einer randomisierten Studie benötigt keine umfangreiche statistische Auswertung[6] und wird daher in der experimentellen wissenschaftlichen Studie häufig angesetzt. Im Kontext der (realen) Forschungsförderung ist dieser Rahmen kaum existent. Die Auswahl von Förderprojekten basiert in der Regel auf Qualitätsbewertungen von Expertinnen und Experten und wird selten per Losverfahren entschieden. Zudem erlaubt auch die bestehende Datenlage nur selten empirische Vergleiche von Kontrollgruppen sowie die Anwendung der in Abschnitt 3.2 beschriebenen Verfahren zur multivariaten Analyse.[7]

Daher verwenden Evaluationsstudien häufig Methoden, die sich Kontrollgruppenvergleichen annähern. Beispielsweise wurden in der Evaluation des „Zentralen Innovationsprogramms Mittelstand (ZIM)" (siehe KMU Austria et al. 2019) Befragungsdaten von geförderten Unternehmen verglichen mit Befragungsdaten nicht geförderter Unternehmen, die anhand mehrerer Merkmale (wie z. B. Unternehmensgröße, FuE-Aktivität, Exportaktivität) den geförderten Unternehmen ähneln. Zudem wird in Evaluationen häufig auf einen Mixed-Methods-Ansatz (Triangulation) aus Online-Befragungen von geförderten bzw. beauftragten Unternehmen, Desktop-Recherchen sowie Interviews mit Expertinnen und Experten zurückgegriffen, um Zielerreichungs-, Wirkungs- und Wirtschaftlichkeitskontrolle mit verschiedenen Indikatoren zu beschreiben (siehe z. B. Bießlich et al. 2019; Stehnken et al. 2021). Dabei werden quantitative Wirkungsanalysen mit verschiedenen Methoden durchgeführt, um valide Ergebnisse zu erzielen, wenn die – nachträgliche – Erhebung von Daten in Kontrollgruppen nicht (oder nur mit enorm hohen Aufwand) möglich ist. Diese Messungen entsprechen jedoch keineswegs kausalen Analysen.

Im Rahmen der genannten Evaluationsstudien wird ein hoher Aufwand betrieben, um valide Ergebnisse zur Wirksamkeit zu erhalten. Nichtsdestotrotz sind Kausalanalysen und Kontrollgruppenvergleiche bislang vergleichsweise selten möglich, da geeignete Daten hierfür nicht vorhanden (oder nicht verwendbar) sind (siehe Abschnitt 4.2). Idealerweise sollten daher Überprüfungen der Wirksamkeit von (Förder-)Maßnahmen schon bei ihrer Konzeption mitgedacht werden. Im folgenden Abschnitt sind Beispiele angeführt, wie die in diesem Papier beschriebenen Methoden stärker in der Forschungsförderung angewendet werden könnten.

5.2 Möglicher Einsatz von RCT-Studien in Förderformaten

Der Einsatz von randomisierten kontrollierten Studien (RCT-Studien) in der Projektförderung ist denkbar. Seit 2017 erprobt die VolkswagenStiftung etwa ein neues Auswahlverfahren für Projektanträge: In ihrer Förderlinie „Experiment!" werden nicht nur Projekte von einer unabhängigen Jury ausgewählt, sondern zusätzlich weitere Vorhaben aus den zum Programmziel passenden und qualitativ uneingeschränkt förderbaren Anträgen ausgelost. Als Begründung für die Anwendung dieses Verfahrens wird u.a. von der VolkswagenStiftung angeführt, dass ein Losentscheid frei von jeglicher Verzerrung (bias) und von Einflüssen durch Gruppendynamik sei und dadurch risikoreiche Forschungsvorhaben stärker begünstigt werden (siehe VolkswagenStiftung 2020).

Im Kontext der Evaluation von Innovations- und Forschungsförderung können solche Ansätze stärker genutzt werden. Dies erfordert allerdings eine zufällige Auswahl einer Gruppe, die eine Förderung erhält und einer Gruppe, die keine Förderung erhält. Dies kann kritisch gesehen werden, da auf eine qualitative „Bestenauswahl" verzichtet wird. Für die Anwendung von Kontrollgruppenanalysen auf diese Auswahl müsste aber nicht gänzlich verzichtet werden: Beispielsweise wäre im Rahmen einer Fördermaßnahme denkbar, dass zunächst von Expertinnen und Experten beurteilt wird, welche Anträge die formalen Förderkriterien sowie ggf. eine Mindestqualität bzgl. des Inhalts erfüllen. Diese Projekte qualifizieren sich dann für ein Losverfahren, in dem zufällig ausgewählt wird, welche Projekte die Förderung erhalten und welche nicht. Ein derartiges Verfahren sollte vor Beginn der Einholung der Förderanträge kommuniziert und begründet werden, um Akzeptanz für diese Methodik bei Antragstellenden zu schaffen (siehe EFI 2024). Ein solches Verfahren würde Kontroll- und Vergleichsgruppenanalysen ermöglichen, die geeignet sind, um kausal den Effekt der Fördermaßnahme zu bestimmen.

6 Prinzipiell wäre ein unabhängiger t-test zum Gruppenvergleich ausreichend, wenn die Randomisierung vollständig gelungen ist.
7 Ferner lassen sich auch die in Abschnitt 2 beschriebenen Effekte, wie der Spillover-Effekt oder der Hawthorne-Effekt, selten gänzlich ausschließen.

Beispiel für Auswahlverfahren mit Komponenten der Qualitätsauswahl und Randomisierung

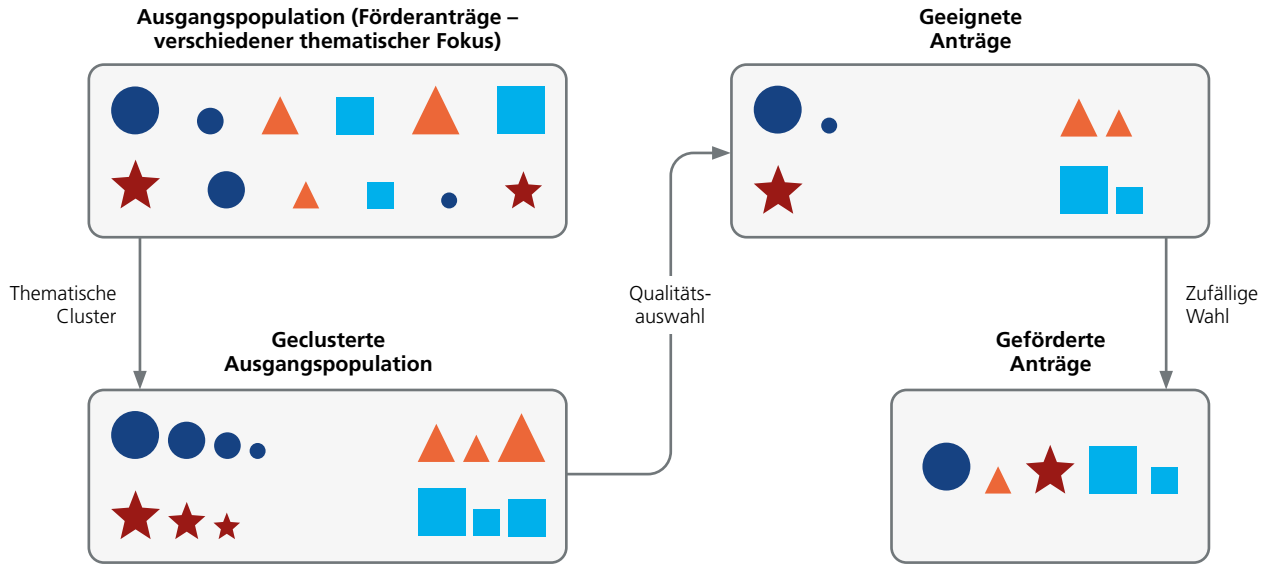


Abbildung 2: Beispiel für Auswahlverfahren mit Komponenten der Qualitätsauswahl und Randomisierung (Quelle: Institut für Innovation und Technik)

Eine wichtige Voraussetzung für diese Analyse ist auch die Formulierung von klaren Zielstellungen einer jeweiligen Maßnahme, die eine klare Wirkungsmessung ermöglicht. Zudem muss gewährleistet sein, dass auch die nicht geförderten Antragstellenden nach Beginn der Förderung weiterhin (in Bezug auf die abhängige, aber idealerweise auch unabhängige Variablen) „gemonitored" werden können. Die politische Durchsetzbarkeit und Akzeptanz dieses Vorgehens bei der Zielgruppe ist ebenfalls zu prüfen.

Angeregt wird, die in diesem Papier dargestellten Verfahren beim Design von Fördermaßnahmen stärker in den Blick zu nehmen. Der verstärkte Einsatz von Kausalanalysen kann dazu beitragen, Maßnahmen mit starker Wirkung entsprechend zu priorisieren und Einsparungspotenziale bei wenig wirksamen Maßnahmen vorzunehmen – und so auch bei angespannten Haushalten eine effiziente politische Förderung zu gewährleisten.

5.3 Kausalanalysen ohne RCT-Verfahren

Kontroll- und Vergleichsgruppen können auch ohne Zufallsentscheid gebildet werden, beispielsweise indem Forschende gezielt bestimmte Gruppen auswählen, die bestimmte Merkmale gemeinsam haben – und sich in anderen Merkmale unterscheiden. Dies kann z. B. durch eine gezielte Auswahl von Teilnehmenden aufgrund ihrer demografischen Merkmale, durch das Ausnutzen von Schwellenwerten oder durch Matching-Verfahren erfolgen (siehe Abschnitt 2.2). Solche Verfahren stellen einige Anforderungen an Datensätze (siehe Abschnitt 4.1.), sodass auch in den Fällen, in denen sich randomisierte Verfahren nicht anbieten, die Analyse der Maßnahmenwirkung bereits bei der Konzeption der (Förder-)Maßnahme mitgedacht werden sollte.

6 Zusammenfassung

Das Wissen darüber, ob und inwiefern Fördermaßnahmen eine kausale Wirkung erzeugen, kann eine wichtige Grundlage evidenzbasierter politischer Entscheidungen darstellen. Evaluationen mit Kontroll- und Vergleichsgruppen können dieses Wissen liefern – setzen allerdings voraus, dass die entsprechende Methodik bereits im Design der Fördermaßnahme mitgedacht wird.

Im vorliegenden Beitrag wurden verschiedene Methoden zur Bildung von Kontroll- und Vergleichsgruppen sowie Methoden zur Analyse kausaler Effekte vorgestellt.

Das Vorgehen erfolgt dabei grundsätzlich in drei Schritten:

1. Bildung von zwei Gruppen: Im ersten Schritt werden aus einer Grundpopulation zwei Gruppen gebildet, deren Zusammensetzung vergleichbar ist. Dabei erfolgt die Auswahl und Zuordnung zu einer der beiden Gruppe in der Regel zufällig. Alternativ können geeignete Matching-Verfahren und/oder – in Ausnahmefällen – natürliche Experimente herangezogen werden. Dadurch soll gewährleistet werden, dass beide Gruppen im statistischen Sinne identisch sind.

2. Treatment in einer Gruppe: Im zweiten Schritt erhält eine der beiden Gruppen ein sogenanntes Treatment, also eine bestimmte „Behandlung", die darauf abzielt, im Idealfall eine messbare Wirkung zu entfalten. Die Kontrollgruppe erhält ein solche Behandlung nicht.

3. Analyse: Im dritten und letzten Schritt werden statistische Tests durchgeführt, mit denen die statistischen Unterschiede zwischen Gruppen analysiert und sichtbar gemacht werden. Dafür reichen in der Regel einfache Inferenz-Statistiken, wie z.B. ein zweiseitiger t-test, aus. Durch die skizzierte Methodik können signifikante Unterschiede zwischen Gruppen kausal auf das Treatment zurückgeführt werden.

Die Aufteilung einer Grundpopulation in möglichst identische Gruppen steht allerdings grundsätzlich im Widerspruch zur Auswahl der „besten" Projekte, die als gängige Praxis im Rahmen von Gutachterverfahren praktiziert wird. Denn: Gutachterentscheide teilen Anträge in gute (geförderte) und weniger gute (nicht geförderte) Projekte ein. Die Gruppen sind damit nicht vergleichbar – und auch nicht zufällig zusammengestellt, sondern unterscheiden sich (mindestens) in der Qualität der Antragstellung. Damit sind Kontroll- und Vergleichsgruppenanalysen weitestgehend ausgeschlossen.

Dennoch lässt sich festhalten, dass nicht vollständig auf qualitätsorientierte Gutachterverfahren verzichtet werden muss, um kontroll- und vergleichsgruppen-basierte Analysen der Fördermaßnahmen durchführen zu können. Denkbar wäre etwa eine „temporäre" Gruppenzuteilung nach Maßgabe der Zufälligkeit – beispielsweise im Rahmen einer Förderlinie pro Jahr. Auch könnte eine zufällige Vergabe unter denjenigen Anträgen in Betracht gezogen werden, die bereits die zuvor spezifizierten Mindestanforderungen grundsätzlich erfüllen.

Insbesondere in Anbetracht immer knapper werdender Haushalte sollten wissenschaftliche Ansätze evidenzbasierter Evaluationsdesigns künftig stärker in den Fokus rücken – vor allem dann, wenn es darum geht, die kausale Wirkung von Fördermaßnahmen verlässlich zu überprüfen.

7 Literatur

Andersen, Simon C.; Beuchert, Louise; Nielsen, Helena S.; Thomsen, Mette K. (2020): The Effect of Teacher`s Aides in the Classroom: Evidence from a Randomized Trial. In: Journal oft he European Economic Association, Jg. 18, Heft 1, S. 469–505.

Arni, Patrick (2012): Kausale Evaluation von Pilotprojekten: Die Nutzung von Randomisierung in der Praxis. In: IZA Standpunkte, Nr. 52, S. 1–27, online verfügbar unter: https://www.econstor.eu/bitstream/10419/91813/1/sp52.pdf, letztes Abrufdatum: 05.02.24.

Austin, Peter C. (2014): A comparison of 12 algorithms for matching on the propensity score. In: Statistics in medicine, Jg. 33, Heft 6, S. 1057–1069.

Baser, Onur (2006): Too much ado about propensity score models? Comparing methods of propensity score matching. In: Value in Health, Jg. 9, Heft 6, S. 377–385.

Bießlich, Susann; von Engelhardt, Sebastian; Kaufmann, Peter; Kerlen, Chirstiane; Kind, Sonja; Kofler, Jakob; Marcher, Anja; Nindl, Elisabeth; Robeck, Martin-Simon; Rodriguez Rivera, Karoline; Zinke, Guido. (2021): Evaluation des nationalen Programms für Weltraum und Innovation, online verfügbar unter: https://repository.fteval.at/id/eprint/640/1/evaluation-des-nationalen-programms-fur-weltraum-und-innovation.pdf, letztes Abrufdatum: 20.03.24.

Bundesministerium der Justiz (BMJ) (2024): Gesetz über die Statistik für Bundeszwecke, online verfügbar unter: https://www.gesetze-im-internet.de/bstatg_1987/__16.html, letztes Abrufdatum: 05.02.24.

Card, David; Krueger, Alan B. (1993): Minimum wages and employment: A case study of the fast food industry in New Jersey and Pennsylvania, NBER Working Paper Nr. 4509, S.1–43, online verfügbar unter: https://www.nber.org/system/files/working_papers/w4509/w4509.pdf, letztes Abrufdatum: 30.01.24.

Europäische Kommission (2023): Staff Working Document on the Synthesis of the Findings of the Evaluations of the European Structural and Investment Fund Programmes 2014–2020, online verfügbar unter: https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:52023SC0022, letztes Abrufdatum: 26.04.2024.

Europäische Kommission (2022): Entwicklung und Beauftragung von kontrafaktischen Wirkungsanalysen – Ein praktischer Leitfaden für ESF-Verwaltungsbehörden, online verfügbar unter: https://op.europa.eu/de/publication-detail/-/publication/dd4a4fc7-42a3-11ec-89db-01aa75ed71a1, letztes Abrufdatum: 26.04.2024.

Expertenkommission für Forschung und Innovation (EFI) (2024): Gutachten zu Forschung, Innovation und technologischer Leistungsfähigkeit Deutschlands – Gutachten 2024, online verfügbar unter: https://www.e-fi.de/fileadmin/Assets/Gutachten/2024/EFI_Gutachten_2024_24124.pdf, letztes Abrufdatum: 06.03.24.

Imbens, Guido W.; Lemieux, Thomas (2008): Regression discontinuity designs: A guide to practice. In: Journal of Econometrics, Jg. 142, Heft 2, S. 615–635.

Kleibergen, Frank; Zivot, Eric (2003): Bayesian and classical approaches to instrumental variable regression. In: Journal of Econometrics, Jg. 114, Heft 1, S. 29–72.

KMU Austria, Institut für höhere Studien – Institute for Advanced Studies, RWK Kompetenzzentrum (2019): Evaluation des Zentralen Innovationsprogramms Mittelstand (ZIM) – Endbericht, online verfügbar unter: https://www.zim.de/ZIM/Redaktion/DE/Publikationen/Studien-Evaluationen/evaluation-zim-2019-07.pdf?__blob=publicationFile&v=1, letztes Abrufdatum: 21.03.24.

Kuß, Alfred; Wildner, Raimund; Kreis, Henning (2014): Marktforschung: Grundlagen der Datenerhebung und Datenanalyse. Springer-Verlag.

Snow, John (1855): On the comparative mortality of large towns and rural districts, and the causes by which it is influenced. Journal of Public Health and Sanitary Review, Jg. 1, Heft 4, S. T16–T24.

Stehnken, Thomas; Astor, Michael; Neumann, Michael; Talamo, Jonathan Aton; Ploder, Michael; Breitfuss-Loidl, Marija; Rosenball, Ricarda; Rammer, Christian; Gottschalk, Sandra; Peters, Bettina (2021): Evaluation des BMBF-Förderprogramms „Photonik Forschung Deutschland – Abschlussbericht, Kurzfassung, online verfügbar unter: https://www.photonikforschung.de/media/Publikationen/A1_Kurzfassung_Evaluation_Photonik-bf.pdf, letztes Abrufdatum: 11.04.24.

VolkswagenStiftung (2020): 8 Thesen für Losentscheid in der Forschungsförderung, online verfügbar unter: https://www.volkswagenstiftung.de/de/news/aktuelles/8-thesen-fuer-ein-loselement-der-forschungsfoerderung, letztes Abrufdatum: 21.03.24.

Wangler, Leo (2014): Evaluation von Forschungs-, Entwicklungs- und Innovationsbeihilfen: Zu einer praktischen Umsetzung von Vergleichsgruppenansätzen, iit-Persepktive 22, online verfügbar unter: https://www.iit-berlin.de/iit-docs/b2a76682a7f5430894dac97ef046df70_iit%20perspektive%2022_Evaluation%20von%20Forschung%20und%20Entwicklung.pdf, letztes Abrufdatum: 07.03.24.

Wangler, Leo (2015): Evaluation von Forschungs-, Entwicklungs- und Innovationsbeihilfen: Zu einer praktischen Umsetzung von Vergleichsgruppenansätzen. In: Zeitschrift für Evaluation, Jg. 14, Heft 1, S. 106–115.

DOI: 10.23776/2024_11