Präv Gesundheitsf
https://doi.org/10.1007/s11553-023-01026-7
Eingegangen: 23. Dezember 2022
Angenommen: 1. März 2023

Sebastian Schneeweiss

Abteilung für Pharmakoepidemiologie und Pharmakoökonomie, Klinik der Inneren Medizin, Brigham und Women's Hospital, Harvard Medical School, Boston, USA

Von Real-World-Daten zur Real-World-Evidenz: eine praktische Anleitung

Die Real World Evidence (RWE) – die Evidenz zu kausalen Behandlungseffekten, die aus elektronischen Daten aus dem Versorgungsalltag gewonnen wird – hat bei politischen Entscheidungsträgern, Kostenträgern und Ärzteschaft große Aufmerksamkeit erlangt. RWE soll die grundlegenden Erkenntnisse über die Wirksamkeit von medizinischen Interventionen, die wir aus randomisierten kontrollierten Studien („randomized controlled trials" [RCT]) gewinnen, ergänzen, indem sie Informationen über die Wirksamkeit in der klinischen Praxis liefert. Anstelle einer Dichotomie von RCT vs. RWE lassen sich bei RCT-Studien zunehmend pragmatische Elemente beobachten. RCT-Studien nähern sich somit allmählich den RWE-Studien an; dieser Trend lässt sich auch umgekehrt beobachten.

Warum haben wir so viel mehr Vertrauen in RCT- als in RWE-Studien? Es gibt zahlreiche Beispiele, in denen RWE-Studien in deutlichem Widerspruch zu RCT standen: Man denke nur an die Hormonersatztherapie bei postmenopausalen Frauen, für die sich später herausstellte, dass eine Hormonersatztherapie mit einem erhöhten Risiko für die koronare Herzkrankheit einherging und nicht etwa, wie zunächst postuliert, zu deren Reduzierung führte [12, 15]. Es wurde angenommen, dass eine Vitamin-E-Supplementierung vor koronarer Herzkrankheit schützt [27, 40], aber dieser Effekt konnte in einem groß angelegten RCT nicht bestätigt werden [50]. Auch die erhebliche Verringerung von Knochenbrüchen und Demenz, die in RWE-Studien mit der Einnahme von Statinen in Verbindung gebracht wurde, konnte in RCT nicht bestätigt werden [3, 13]. Die Pharmakoepidemiologie hat in den letzten zwei Dekaden erheblich Fortschritte gemacht und Techniken entwickelt, die die Validität von RWE substantiell erhöhen und Situationen identifiziert, in denen RWE unzuverlässig bleibt.

Unabhängig davon, wie Evidenz generiert wird, muss diese, um handlungsrelevant zu sein, intern valide und auf eine definierte Zielpopulation generalisierbar sein.

Real-World-Daten und ihre Verwendung in der Forschung

Typische Real-World-Datenquellen

Moderne Gesundheitssysteme generieren eine Fülle elektronisch gespeicherter Informationen zu einzelnen Patienten. Diese fortlaufenden Datenströme können über eindeutige Personenidentifikatoren longitudinal miteinander verknüpft werden. Im Gegensatz zu stark kontrollierten Forschungsdaten spiegeln diese Daten die klinische Praxis wider (**Abb. 1**). Viele RWE-Studien verwenden heutzutage solche Längsschnittdaten zu medizinischen Produkten, Interventionen und gesundheitsrelevanten Ereignissen, die im Versorgungsalltag erfasst werden. Warum solche Daten in der RWE-Forschung an Beliebtheit gewonnen haben, lässt sich folgendermaßen begründen:

1. Sie weisen meist eine höhere Repräsentativität auf, als die meisten experimentellen Studien.
2. Sie erfassen medizinische Eingriffe und den Arzneimitteleinsatz prospektiv, in großer Ausführlichkeit (USA, UK) und sind dabei weder auf eine Einverständniserklärung noch das Erinnerungsvermögen der Patienten angewiesen.
3. Sie erfordern keine Experimente am Menschen und sind schneller sowie kostengünstiger als die meisten klinischen Studien oder andere Studien, die auf der Erhebung von Primärdaten beruhen.
4. Die prospektive, longitudinale Erfassung von Versorgungskontakten mit dem Gesundheitssystem inklusive der Dokumentation des Datums, zu dem eine Leistung erbracht wurde, schafft Klarheit über den zeitlichen Ablauf der Versorgung. Dies stellt eine Voraussetzung für kausale Schlüsse hinsichtlich der Wirksamkeit einer Behandlung dar.

Über die Möglichkeiten und Grenzen verschiedener Datentypen ist bereits viel geschrieben worden [11, 33]. **Tab. 1** führt die wesentlichen Aspekte auf.

Von Real-World-Daten zur RWE

Es ist unabdingbar, eine Datenquelle vollständig zu verstehen, ehe man versucht, damit Evidenz zu kausalen Behandlungseffekten zu generieren.

Der Prozess der Planung, Implementierung und Auswertung einer RWE-Studie umfasst 3 Ebenen, die gemeinsam einen sequenziellen Arbeitsablauf bilden (**Abb. 2**):

Published online: 12 April 2023

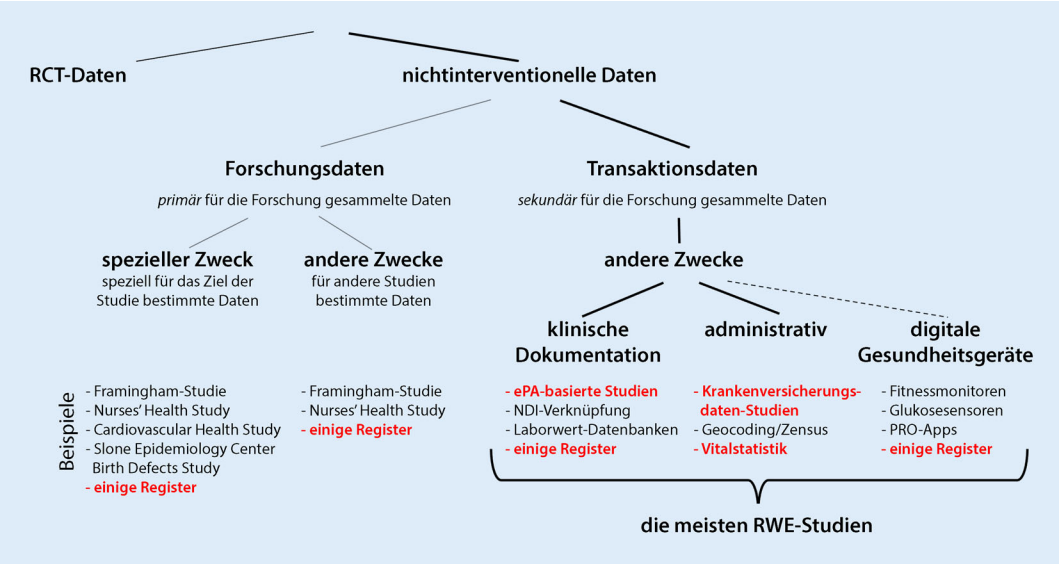

Abb. 1 ◀ Eine Taxonomie der Daten aus dem Gesundheitswesen, die oftmals für RWE-Studien verwendet werden. *RCT* randomisierte kontrollierte Studie, *ePA* elektronische Patientenakte, *NDI* „national death index", *PRO* „patient-reported outcomes", *RWE* Real-World-Evidenz

1. Auf der Designebene wird das Studiendesign präzisiert. Hierzu stellt man sich am besten die randomisierte Studie vor, die man idealerweise durchführen würde und mit Real-World-Daten nachbilden möchte (den sog. „target trial"). Dies führt häufig zu einem „New-user active-comparator"-Kohortendesign [17, 26, 30], welches sich bei der Vorhersage und Replikation von Ergebnissen aus RCT bewährt hat [6, 24, 25].
2. Auf der Messebene wird der longitudinale elektronische Datenstrom auf Patientenebene in Variablen umgewandelt. Anhand dieser Variablen lassen sich die Studienpopulation, der Gesundheitszustand vor der Behandlungsexposition (zur Kontrolle von Confoundern bei fehlender Randomisierung zu Baseline), der Behandlungsstatus sowie behandlungsbedingte Outcomes identifizieren.
3. Die Analyseebene befasst sich mit der Schätzung des kausalen Behandlungseffekts unter Berücksichtigung des Mechanismus zur Datenerhebung. Aufgrund ihrer Eignung für große sekundäre Datenbanken haben Propensity-Score-Analysen zur Ausbalancierung von Patientenmerkmalen zwischen unterschiedlichen Behandlungsgruppen an Popularität gewonnen [46]. Mit zusätzlichen Methoden können systematische Verzerrungen und differentielles Follow-up verringert werden. Zudem können weitere bekannte Biasarten wie der „immortal time bias", eine Adjustierung kausaler Mediatoren oder umgekehrte Kausalität vermieden werden.

In den folgenden Abschnitten werden diese drei Ebenen erläutert.

Wahl des Studiendesigns

Grundlegende Überlegungen bei der Auswahl des RWE-Studiendesigns

Die klinische Fragestellung bestimmt die Wahl des Studiendesigns. In den meisten RWE-Studien wird die Wahl des Studiendesigns zusätzlich durch den Inhalt und die Limitationen der zugrunde liegenden Datenquellen beeinflusst. Im Rahmen eines hypothetischen kontrafaktischen Experiments würde ein Patient behandelt werden und anschließend ein Eintreten oder Nichteintreten des gesundheitsbezogenen Outcomes beobachtet werden. Anschließend würde der Vorgang wiederholt werden, indem die Zeit kontrafaktisch zurückgedreht würde, der Patient jedoch unbehandelt bleiben würde bei gleichzeitiger Konstanthaltung aller anderen Faktoren (kontrafaktische Erfahrung). Dieses hypothetische Experiment würde den kausalen Zusammenhang bei diesem Patienten nachweisen.

Nachdem die Zeit in unserer Realität nicht zurückgedreht werden kann, lassen wir sie weiterlaufen und beobachten, wie Patienten episodisch exponiert und dann wieder nicht exponiert sind. Das resultierende „case-crossover design" oder eine „self-controlled case series" kann in Betracht gezogen werden, wenn sich der Expositionsstatus eines Patienten im Zeitverlauf ändert (**Abb. 3**; [21]); wenn beispielsweise bei einem Kopfschmerzmedikament, das angenommen eine kurze Wirkungsdauer hat, das interessierende Ereignis schnell eintritt (z. B. Lebertoxizität). Die meisten RWE-Studien nutzen natürlich auftretende Behandlungsunterschiede zwischen Patienten und verwenden daher ein Kohortenstudiendesign mit zeitgleicher Kontrollgruppe. Wenn die Datenerfassung zeitaufwändig oder teuer ist, können innerhalb von Kohorten effiziente Stichprobendesigns wie Case-control-, Case-cohort- oder zweistufige Stichproben verwendet werden [30]. Unterschiede in der Arzneimittelbehandlung zwischen Patientengruppen oder übergeordneten Entitäten (d. h. zwischen Ärzten, Krankenhäusern, Krankenversicherungen, Regionen usw.) können anhand der Instrumentvariablenschätzung erschlossen werden [1].

Die Auswahl der Vergleichsgruppe ist eine grundlegende Designentscheidung, die die klinische Interpretation erheblich beeinflusst und die Effektgröße stark verändern kann. Das Vergleichspräpa-

rat muss im klinischen Kontext relevant sein und eine praktikable Alternative zum untersuchenden Arzneimittel darstellen. Idealerweise sollte sich die Vergleichspopulation auf Patienten beschränken, die in der klinischen Praxis die gleiche Indikation aufweisen wie die Anwender des untersuchten Arzneimittels [49].

Kohortenstudien und die „Target-trial"-Denkweise zur Vermeidung von Bias

Es ist ein wertvoller didaktischer Beginn jeder RWE-Studienplanung, sich eine randomisierte Studie vorzustellen, die man durchführen würde, wenn dies logistisch und ethisch möglich wäre, und diesen „target trial" dann zu emulieren, d. h. im Design einer RWE-Studien nachzuahmen [14]. Selbst ohne Randomisierung werden so systematische Verzerrungen durch das Studiendesign reduziert und das Design wird klarer. Diese „Target-trial"-Denkweise schafft Klarheit darüber, wann Patiencharakteristika, Exposition und Outcome im Verhältnis zum Studieneintritt gemessen werden sollten. Dies ist entscheidend, um kausale Schlüsse ziehen zu können. Es wird zudem die analytische Strategie einer „as-started"- (auch Intention-to-treat-Analyse genannt) oder einer „As-treated"-Analyse verdeutlicht. Sobald ein „target trial" konzipiert ist, werden durch das Design der Trial-emulierenden RWE-Studie sowie die möglichen Abweichungen vom Trail potenzielle Schwächen der Datenqualität, Datenvollständigkeit und der kausalen Inferenz aufgezeigt [16].

Das Design von Trial-emulierenden RWE-Studien zeigt häufig das Spannungsverhältnis zwischen dem Ziel, hochgradig generalisierbare Ergebnisse zu generieren und der Einschränkungen, die zur Sicherstellung einer hohen Ergebnisvalidität und kausaler Schlussfolgerungen notwendig sind. Es hilft jedoch bei der Identifizierung und Vermeidung von Designschwächen, wie z. B. dem „immortal time bias", der Adjustierung von kausalen Mediatoren und der umgekehrten Kausalität, sowie dem Umgang mit zeitlich variierenden Risikofaktoren und der vorzeitigen Entfernung von anfälligen Patienten [16].

Präv Gesundheitsf https://doi.org/10.1007/s11553-023-01026-7

S. Schneeweiss

Von Real-World-Daten zur Real-World-Evidenz: eine praktische Anleitung

Zusammenfassung

Hintergrund. Die Real World Evidence (RWE) liefert Erkenntnisse über die Wirksamkeit von Behandlungen im klinischen Versorgungsalltag auf Basis von patientenindividuellen Längsschnittdaten, die im Routinebetrieb des Gesundheitssystems anfallen und ergänzt Erkenntnisse aus randomisierten kontrollierten Studien.
Fragestellung. Das Ziel dieses Beitrags ist es, aufzuzeigen, wie robuste und handlungsrelevante RWE generiert werden.
Material und Methode. Bewährte Methoden wurden in einem umfassenden und dennoch prägnanten Überblick zusammengefasst.
Ergebnisse. Die Durchführung von RWE-Studien folgt einem strukturierten Ansatz: (1) Die Designebene verbindet die Forschungsfrage mit dem geeigneten Studiendesign, (2) die Messebene wandelt den longitudinalen Datenstrom auf Patientenebene in Variablen um, der die Studienpopulation, die Patientenmerkmale vor der Exposition, die Behandlung und die Ergebnisse identifiziert, und (3) eine Analyseebene konzentriert sich auf die Schätzung der kausalen Behandlungseffekte.
Schlussfolgerung. Sorgfältig geplante und durchgeführte RWE-Studien ergänzen den Erkenntnisgewinn von randomisierten Studien.

Schlüsselwörter

Real World Evidence · Datenbanken aus dem Gesundheitswesen · Kausale Behandlungseffekte · Confounding · Bias · Epidemiologie

Turning real-world data into real-world evidence: some practical guidance

Abstract

Background. Real-world evidence (RWE), the understanding of treatment effectiveness in clinical practice generated from longitudinal patient-level data that the routine operation of the healthcare system produces, complements evidence from randomized, controlled trials (RCTs).
Objectives. The aim of this contribution is to highlight how robust and actionable RWE is generated.
Methods. Best practices were condensed into a comprehensive yet concise guide.
Results. Conducting RWE studies follows a structured approach: (1) A design layer connects the research question with the appropriate study design, (2) a measurement layer transforms the longitudinal patient-level data stream into variables that identify the study population, the pre-exposure patient characteristics, the treatment, and outcomes, and (3) an analysis layer focuses on the causal treatment effect estimation.
Conclusions. RWE, if properly conducted, is a useful complement to randomized trial evidence.

Keywords

Real-world evidence · Healthcare databases · Causal treatment effects · Confounding · Bias · Epidemiology

„New-user"-Kohorte

Die Untersuchung von Personen, die eine medikamentöse Behandlung oder medizinische Intervention neu begonnen haben, hat mehrere Vorteile, insbesondere, wenn diese mit Personen verglichen werden, die eine praktikable Behandlungsalternative neu begonnen haben. Weil die Patienten in beiden Gruppen mit den entsprechenden Behandlungen neu begonnen haben („new users"), wurden sie folglich zuvor von einem Arzt untersucht, der entschieden hat, dass sie von der entsprechenden Behandlung profitieren würden. Diese Tatsache führt zu vergleichbaren Behandlungsgruppen, die sich in Bezug auf Merkmale ähneln, welche in einer gegebenen Datenquelle sowohl beobachtbar als auch nicht beobachtbar sein können [26].

Durch die klare zeitliche Abfolge, in der Confounder vor Beginn der Behandlung erfasst werden, wird der Fehler vermieden, für Behandlungsfolgen (kausale Mediatoren) zu adjustieren. Aufgrund des genau definierten Startpunkts bei

Tab. 1 Häufig verwendete Datenquellen aus der klinischen Praxis und einige ihrer Merkmale

Datenquelle	Vorzüge	Überlegungen
Abrechnungsdaten der Krankenkassen		
Abrechnungsdaten von Krankenkassen umfassen longitudinale Daten auf Patientenebene aller Kontakte mit dem professionellen Gesundheitssystem, einschließlich ärztlicher Leistungen und Krankenhausaufenthalte, dazugehörige Diagnosen und Verfahren sowie alle ambulanten Arzneimittelverordnungen. Daneben sind grundlegende demografische und versicherungsbezogene Informationen enthalten. Die Daten enthalten die Abrechnungscodes, die von den Leistungserbringern an die Kostenträger übermittelt werden, z. B. private Krankenkassen, gesetzliche Krankenkassen, AOK	In Abrechnungsdaten werden oft sehr große Populationen erfasst. Sie bieten eine vollständige longitudinale Aufzeichnung aller medizinischen Kontakte, im Gegensatz zu elektronischen Patientenakten, bei denen es zu Datenverlusten durch Leistungen, die außerhalb eines Netzwerks erbracht wurden. Krankenversicherungsdaten variieren weltweit hinsichtlich der Aussagekraft für die Gesamtbevölkerung, des Umfangs und der Tiefe der enthaltenen Informationen, der Datenqualität und -vollständigkeit sowie der Verknüpfbarkeit mit anderen Datenquellen, z. B. Vitalstatistiken, Krebsregistern, elektronischen Patientenakten und Laborbefunden. Mehrere nationale Gesundheitssysteme, beispielsweise jene in skandinavischen Ländern, haben eine universelle, lebenslange Gesundheitsversorgung	Die Datenbanken von Versicherungen mit hoher Mitgliederfluktuation sind für Untersuchung längerfristiger Outcomes eher weniger geeignet, da hier die longitudinale Nachverfolgung begrenzt ist. Da es sich bei Abrechnungsdaten um Transaktionsdaten handelt, die eher zu administrativen als zu Forschungszwecken erhoben werden, müssen Forschende genau prüfen, ob die Messwerte zentraler Variablen für die Beantwortung einer bestimmten Studienfrage angemessen sind. Beispielsweise werden Informationen über klinisch relevante Parameter wie Body Mass Index (BMI), Ernährungsgewohnheiten, Familienanamnese und bestimmte Lebensstilfaktoren wie Rauchen und Alkoholkonsum in den Abrechnungsdaten nicht zuverlässig erfasst oder erfordern die zusätzliche Verknüpfung mit anderen Datenquellen
Daten aus der elektronischen Patientenakte (ePA)		
Die ePA dient der klinischen Dokumentation und enthält ein großes Spektrum an gesundheitsrelevanten Information. Die Daten stammen von unterschiedlichen Ärzten, die an der Versorgung des Patienten beteiligt sind und bestehen sowohl aus vorgegebenen als auch aus Freitextdaten	Eine ePA enthält eine Fülle informativer, klinischer Daten, die in den Abrechnungsdaten der Krankenversicherungen fehlen können, aber notwendig sind, um bestimmte Studienfragen beantworten zu können. Dies schließt Informationen über Symptome, Untersuchungsergebnisse, Laborbefunde und Verfahren, Diagnosen und Behandlungspläne sowie Anamnese und sozialer Hintergrund mit ein. ePA werden zunehmend in der Forschung verwendet und dienen als Quelle für bestimmte klinische Informationen, die in Abrechnungsdaten häufig fehlen. So können beispielsweise Hb_{A1c}-Werte, die Dauer eines Diabetes, BMI, Ernährungskontrolle sowie Arztberichte aus den Daten der ePA extrahiert werden	Eine wesentliche Einschränkung der meisten ePA-Daten besteht darin, dass nur diejenigen Patienteninformationen dokumentiert werden, die von Leistungserbringern erzeugt werden, die das System pflegen. Das daraus resultierende „Datenleck" ergibt ein unvollständiges Bild, wenn die Versorgung des Patienten lückenhaft ist und entsprechende Informationen unter den Leistungserbringern nicht elektronisch ausgetauscht wird. Da der ursprüngliche Zweck der ePA nicht die Forschung, sondern die Unterstützung der klinischen Versorgung ist, ist Unvollständigkeit eine häufig auftretende Hürde bei der Sicherstellung der Datenqualität. Darüber hinaus gibt es keinen allgemein anerkannten Standard für die Art der Daten, die in ePA enthalten sein sollten, was zu einer Heterogenität bei den Aufzeichnungen führt
Patientenregister		
Bei einem Patientenregister werden nicht-interventionelle Studienmethoden verwendet, um systematisch longitudinale Informationen über Patienten mit einer bestimmten Erkrankung oder Behandlung zu erheben	Im Gegensatz zu Abrechnungsdaten und Daten aus der ePA, können Register sehr spezifische und detaillierte klinische Informationen mit hoher Datenvollständigkeit erfassen, die für manche RWE-Studien notwendig sind. Beispielsweise Informationen über diagnostische Tests, prognostische Tests, angebotene und erhaltene Therapien, Familienanamnese der Erkrankung, verhaltens- und umweltbedingte Risikofaktoren, Symptome und Krankheitsverlauf sowie grundlegende demographische Informationen wie Alter und Geschlecht	Je nach Register kann es sich um ein höchst selektives Patientensegment handeln und es wird möglicherweise nicht die routinemäßige Versorgung widergespiegelt. Die longitudinale Dokumentation von Medikamenteneinnahmen wird häufig rückwirkend über Patientenbefragung erhoben, was eigene Einschränkungen hat
Patientengenerierte Daten		
Hierbei handelt es sich beispielsweise um Daten aus Umfragen, Fragebögen, Smartphone-Apps und sozialen Medien, die eine kontinuierliche Datenerfassung ermöglichen. Die Informationen werden eher von den Patienten als von den Anbietern bereitgestellt	Fragebogen-/Umfragedaten liefern Messwerte zur Lebensqualität, die in anderen Datenquellen schwer zu finden sind. Dies kann für die Pharmakovigilanz von besonderer Bedeutung sein, insbesondere für seltene unerwünschte Ereignisse im Zusammenhang mit Behandlungen und für Faktoren, die die Therapieadhärenz, Verhaltensweisen und Einstellungen der Patienten vorhersagen. Einige Daten ermöglichen Aufzeichnungen in Echtzeit, das die Beobachtung ausgewählter Messwerte und Symptome ermöglicht	Die Verwendung dieser Quellen impliziert ein Vertrauen in selbstberichtete Variablen, was zu Recall-Bias, selektives Berichten („reporting bias") und fehlenden Daten bei wichtigen Patientenmerkmalen führen kann. Begrenzte Generalisierbarkeit durch selektierte Teilnehmergruppen und begrenzte interne Validität, da die berichteten klinischen Ergebnisse oft nicht validiert sind und die Authentizität oft nicht nachweisbar ist. Die Daten sind daher nur in bestimmten Bereichen nach sorgfältiger Bewertung und Prüfung nützlich

Tab. 1 (Fortsetzung)

Datenquelle	Vorzüge	Überlegungen
Vitalstatistiken		
Vitalstatistiken enthalten Informationen über Todesdatum und in einigen Fällen auch über die Todesursache aus der ärztlich ausgestellten Sterbeurkunde	Präzise und vollständige Informationen zum Todesdatum sind für viele Studien mit klinischen Outcomes von entscheidender Bedeutung	Angaben zur Todesursache sind oft weniger zuverlässig und werden in der Regel anhand von Oberbegriffen angegeben, z. B. kardiovaskuläre Todesfälle, Todesfälle aufgrund von malignen Erkrankungen, Verletzungen (einschließlich Selbstverletzungen), Atemwegserkrankungen usw.
Verknüpfung unterschiedlicher Datenquellen		
Daten aus zwei oder mehreren Quellen werden miteinander verknüpft, um benötigte Informationen zusammenzuführen, vorausgesetzt, es sind angemessene Datenschutzvorkehrungen getroffen worden	Zusammenführung von Daten aus unterschiedlichen Quellen, die die Erfassung umfassender Informationen, die in einem bestimmten Studiensetting benötigt werden, ermöglicht. Beispielsweise kann durch die Verknüpfung von GKV-Abrechnungsdaten mit Daten aus der ePA eine Kombination longitudinaler Nachverfolgung und Informationen zu Kosten (in der ePA fehlend) mit klinischen Variablen (in GKV-Abrechnungsdaten unvollständig enthalten) erreicht werden	Die Validität der Ergebnisse hängt von der Qualität der Datenverknüpfung ab, die oft durch Tokenisierung erfolgt. Es ist teuer, Daten zu verknüpfen und die verknüpften Datenquellen zu warten. Herausforderungen bei der Verknüpfung von Daten ergeben sich aus den unterschiedlichen Zwecken der Datenerhebung, Diskrepanzen bei der Datenaufzeichnung, sowie rechtliche und Aspekte des Datenschutzes

„New-user"-Kohorten lässt sich zudem beurteilen, wie Risiken mit der Dauer der Behandlung variieren. Weil bei der „New-user"-Kohortenstudie das Design eines RCT im Parallelgruppendesign nachgeahmt wird, ist dieses Design für Laien leichter nachvollziehbar. In Zeiten, in der Entscheidungsträger nicht-interventionelle Studien mit erhöhter Skepsis betrachten, oft weil diese undurchdringlich erscheinen und damit die Validität der Studien schwierig zu bewerten ist, sollte dieser Vorteil nicht unterschätzt werden [22]. Beispiele solcher „New-user"-Kohortenstudien umfassen Untersuchungen zum Psychoserisiko bei Kindern und Jugendlichen, die eine Behandlung mit Stimulanzien beginnen [23] sowie zur Wirkung von Statinen auf eine Reihe von gesundheitlichen Outcomes [39].

„Active comparator"

Vergleichende Studien, die mit dem Ziel durchgeführt werden, placebokontrollierte RCT nachzuahmen, leiden häufig unter starken Selektionseffekten bei der Behandlung. Das heißt Personen, die eine Behandlung erhalten, unterscheiden sich von denjenigen Personen, die keine Behandlung erhalten, in einer Weise, die sich analytisch nur schwer vollständig erfassen und kontrollieren lässt. Starkes Confounding tritt auch auf, wenn zwei verschiedene Behandlungsmethoden verglichen werden, z. B. eine medikamentöse Behandlung gegen ein implantierbares Gerät [37]. Der Vergleich wird dadurch verzerrt, dass die gebrechlichsten Patienten sich aufgrund der Risiken nicht operieren lassen werden, obwohl diese Patienten das höchste Risiko für das gewünschte Outcome haben, was den Vergleich verzerrt. Ein Beispiel für ein erfolgreiches „New-user-active-comparator"-Design war die RWE-Studie, welche die Ergebnisse des laufenden CAROLINA-Trials bereits Monate vor dessen Abschluss vorhersagte [25].

Überlegungen zur Variablenerfassung bei der Verwendung von Sekundärdaten aus dem Gesundheitswesen

Es ist viel über Datenstandardisierung und darüber, wie die Qualität von Daten verbessert werden kann, geschrieben worden. Letztendlich laufen alle Diskussionen über die Datenqualität auf die gleiche Frage hinaus: Eignen sich die Daten zur Beantwortung dieser spezifischen Forschungsfrage? In einem doppelt randomisierten Experiment konnte gezeigt werden, dass nicht-randomisierte Studien genau wie RCT unverzerrte Schätzungen liefern können, wenn die Variablen in der zugrunde liegenden Datenbasis ausreichend abgebildet werden [38]. Selbst in einem bestimmten Therapiegebiet kann keine einzelne Datenquelle oder Standardisierungsmethode alle Fragen beantworten. Es kommt darauf an, wie die in **Abb. 2** aufgeführten Expositionen, Outcomes und Confounder erfasst werden. Epidemiologische Prinzipien legen fest, welche erfassten Merkmale benötigt werden. Diese werden in Real-World-Daten nahezu niemals direkt erfasst. Darüber hinaus besteht wenig Einigkeit darüber, welcher Messgenauigkeit ausreicht (**Tab. 2**).

Bei RCT treten ähnliche Probleme auf, wobei viel Zeit und Geld aufgewendet wird um die Vollständigkeit, Genauigkeit, und Rechtzeitigkeit der Messungen zu optimieren. Es mag unmöglich sein, ganze Datenbanken für die Generierung von RWE als geeignet zu zertifizieren, aber wir können zumindest den Prozess der Datengenerierung und -kuration bis zu dem Punkt hin beleuchten, an dem die Daten für eine bestimmte Analyse verwendet werden. Dadurch kann eine Bewertung der erfassten Charakeristika erfolgen, welche wiederum die Grundlage für quantitative Biasanalysen sind [19].

Identifikation der Studienpopulation

Zur Einordnung der Generalisierbarkeit der Ergebnisse ist eine eindeutige Identifikation der Studienpopulation wichtig. Bei RWE-Studien zur Behandlung von Diabetes beginnt der Einschluss von Patienten in die Kohorten meistens mit der zu vergleichenden Behandlung. Kohortenausschlüsse erfolgen anschließend basierend auf der gewünschten

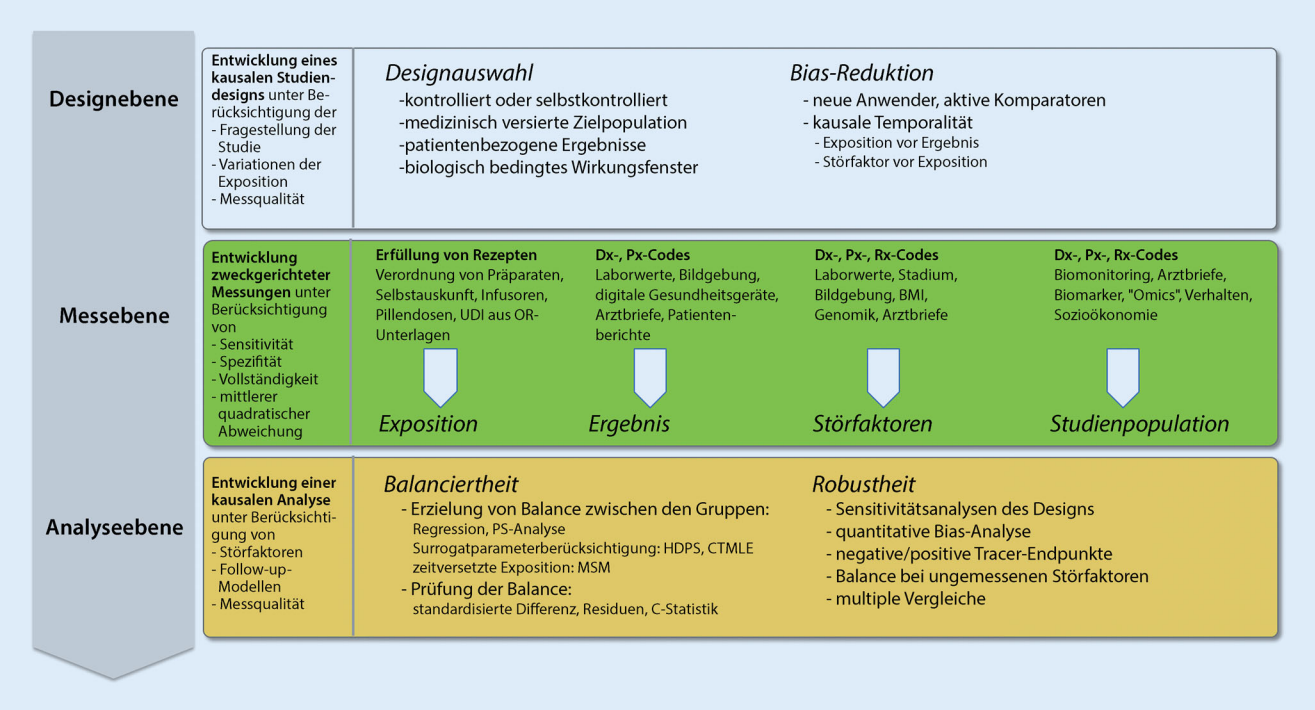

Abb. 2 ▲ Von Real-World-Daten zur Real-World-Evidenz. *UDI* „unique device identifier", *OR* Operationssaal („operation room"), *Dx* Diagnose, *Px* Prozeduren, *BMI* Body Mass Index, *PS* Propensity Score, *CTMLE* „collaborative targeted maximum likelihood estimation", *HDPS* „High-dimensional Propensity Score", *MSM* marginal structural models

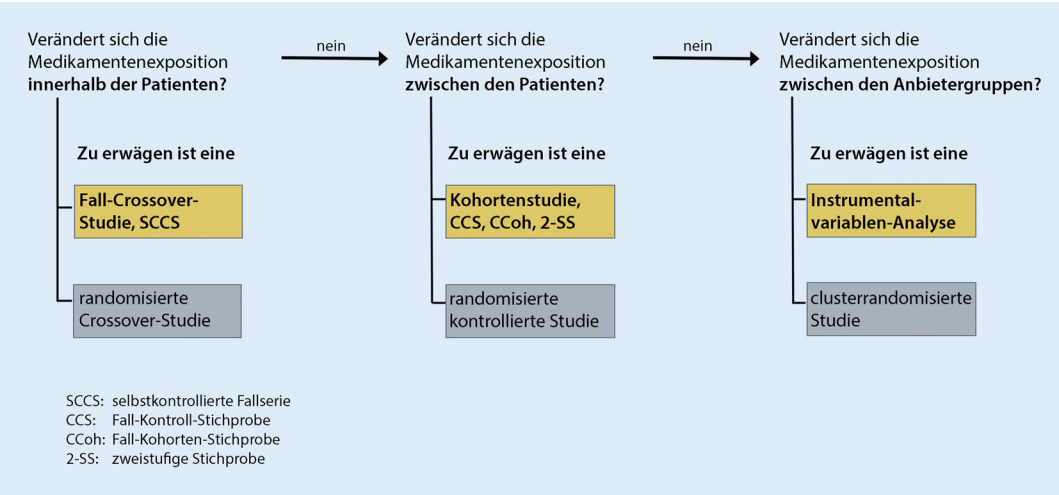

Abb. 3 ◀ Das Studienziel und die Unterschiede in der Therapieexposition bestimmen die Wahl des Designs. *SCCS* selbstkontrollierte Fallserie, *CCS* Fall-Kontroll-Stichprobe, *CCoh* Fall-Kohorten-Stichprobe, *2-SS* zweistufige Stichprobe. (Adaptiert nach Schneeweiss [30])

Altersspanne, des Vorhandenseins bzw. der Abwesenheit bestimmter Diagnosen und dem Vorhandensein von Verlaufsindikatoren. Drei typische Indikatoren bei dem Verlauf des Diabetes sind die Hb_{A1c}-Messung, die Dauer des Diabetes und der Body Mass Index (BMI). Wenn diese Parameter nun für die Interpretation der Ergebnisse wirklich entscheidend wären, würde man eine Datenquelle wählen, in der diese erfasst werden. Häufig sind jedoch eben diese Parameter in solchen Datenquellen nicht enthalten, die groß genug sind, um aussagekräftige Rückschlüsse auf Behandlungseffekte zu ermöglichen. Es ist dann am Forschenden zu entscheiden, ob die Tatsache, dass Patienten mit einer bestimmten Behandlung begonnen haben, ausreicht, um die Patientenpopulation korrekt gemäß dem Krankheitsschweregrad zu kategorisieren, oder ob Zeit und Ressourcen aufgewendet werden sollen, um die nicht in der Datenbasis enthaltenen Parameter zu erheben. Bei RWE-Studien stehen Forschende häufig vor der Herausforderung, zwei suboptimale Optionen abwägen zu müssen, wovon dennoch eine für den Zweck einer gegebenen Studie möglicherweise besser geeignet sein mag.

Behandlungsexposition

Für RWE-Studien ist die Festlegung des Start- und Endpunktes einer interessie-

Tab. 2 Messparameter, die die Aussagekraft von RWE(Real World Evidence)-Studien bestimmen und häufig zitierte Indikatoren für die Datenqualität sind

Studienmerkmale:	Beispiele zur Verbesserung der Messparameter	Typische Näherungswerte für die Datenqualität von Sekundärdaten	Relevante Messcharakteristiken[a]
1) Studienpopulation, Subgruppen	Erfordere zwei Diagnosecodes, um die Spezifität der zugrunde liegenden Erkrankung zu erhöhen	Frühere Erfahrung mit einer Datenquelle, Publikationen Verfügbarkeit von Validierungsstudien Detaillierte Dokumentation der Datenerhebungsmethode Detaillierte Beschreibung des Datenkuratierungsprozesses Detaillierte Beschreibung der Zuordnung zu medizinischer Konstrukten (falls zutreffend) Dokumentation von Veränderungen in der Kodierung im Zeitverlauf	*Binäre Daten, z. B. vorhandene diagnostische Codes*: Sensitivität, Spezifität, positiv-prädiktiver Wert. *Kontinuierliche Daten, z. B. Laborbefunde*: %-Wert fehlt, mittlere quadratische Abweichung. *Zeit bis zum Ereignis („time-to-event")*: Genauigkeit des Auftretens
2) Expositionsmessung	Verwende Abgabedaten anstelle von Verschreibungsdaten um die Vollständigkeit zu erhöhen		
3) Outcome-Messung	Verwende schwerwiegende Ereignisse, z. B. Krankenhauseinweisungen um die Spezifität der Outcome Messung zu erhöhen		
4) Confounder-Messung	Überprüfe ein breites Spektrums potenzieller Confounder und ihrer Proxies zur Einschränkung unbeobachteten Confoundings		

[a]Diese Messcharakteristiken sind für die Quantifizierung eines potenziellen Bias relevant und bewerten die Wahrscheinlichkeit einer kausalen Beziehung zwischen Medikament und Outcome
Adaptiert nach Franklin et al. [8]

renden Behandlung von zentraler Bedeutung. Das Apothekenabgabedatum, welches die Einschränkung von mangelnder Präzision von Patientenangaben („recall bias") umgeht, gilt bei der Festlegung des Beginns einer Arzneimittelexposition als zuverlässig [47]. Dies liegt daran, dass Apotheker Rezepte nur mit wenig Interpretationsspielraum erfüllen und die Erstattung durch die Krankenkasse auf Grundlage detaillierter, vollständiger und genauer elektronisch vorgelegten Abrechnungsdaten erfolgt.

Outcomes

Da in Datenbanken häufig keine detaillierten klinischen Informationen enthalten sind, müssen Forschende die möglichen Auswirkungen einer Missklassifikation des interessierenden Outcomes in Betracht ziehen. Im Allgemeinen ist ein Mangel an Spezifität als schlimmer zu werten als ein Mangel an Sensitivität. Die Schätzung des relativen Risikos wird durch eine Missklassifikation des Outcomes nicht verzerrt, wenn alle Personen ohne das Outcome korrekt als solches identifiziert werden, d. h. 100 %ige Spezifität, selbst wenn wesentlich weniger als 100 % der Patienten mit dem Outcome als solches identifiziert werden, d. h. die Sensitivität ist wesentlich geringer als 100 %, solange die Missklassifikation nicht differentiell ist, d. h. beide Gruppen sind von der Missklassifikation in gleichem Maße betroffen [28]. Studien, in welchen die Missklassifikation von Diagnosen in Abrechnungsdaten untersucht und die Krankenakten zur Bewertung als „Goldstandard" verwendet wurde, haben ergeben, dass die Sensitivität der in Abrechnungsdaten dokumentierten Diagnosen häufig mäßig ist, ihre Spezifität hingegen sehr hoch ist [48]. Dieses Muster ergibt sich daraus, dass, wenn eine Diagnose dokumentiert, kodiert und übermittelt wurde, es sehr wahrscheinlich ist, dass diese Diagnose auch tatsächlich gestellt wurde, insbesondere in stationären Entlassungsdaten [9].

Confounder

Mögliche Störvariablen werden vor Beginn der zu untersuchenden Behandlung erfasst, um zu vermeiden, dass Variablen adjustiert werden, die die Folgen der Behandlung sind und sich kausal auf das interessierende Outcome auswirken, d. h. „causal intermediates" oder Mediatoren [44]. Als Beispiel kann der Vergleich von zwei Antidiabetika dienen, von denen eines den Blutdruck senkt, das andere nicht. In einer Studie zur Inzidenz des Herzinfarkts würde eine Adjustierung des Blutdruckes, der *nach* dem Behandlungsbeginn gemessen wurde, den beobachteten Effekt des Medikaments fälschlich reduzieren, da ein erhöhter Blutdruck ein Mediator eines Herzinfarkts ist. Dahingegen ist eine Adjustierung des Blutdruckes, der *vor* dem Behandlungsbeginn gemessen wurde, notwendig.

Eine Herausforderung von Sekundärdaten aus dem Gesundheitswesen stellt die vollständige und akkurate Erfassung wichtiger Outcome-Prädiktoren dar. Missklassifikationen oder unbeobachtbare Confounder können zu einem residualen Confounding führen, welches im Abschnitt zur Datenanalyse thematisiert wird.

Fehlende Messwerte

Fehlende Messwerte betreffen alle in den vorangehenden Abschnitten thematisierten Aspekte der Variablenerfassung. Wenn für das Untersuchungsziel wesentliche Daten gar nicht oder nur mit erheblicher Lückenhaftigkeit oder Missklassifikation erfasst wurden, dann ist die entsprechende Datenquelle ungeeignet [8, 32]. Im Allgemeinen führen Abrechnungsdaten, die Informationen über Diagnosen und Verfahren enthalten, zu einer Missklassifikation von Informationen und nicht zu fehlenden Messwerten. Ein Nichtvorhandensein eines Diagnosekodes wird in der Regel mit dem Nichtvorhandensein der entsprechenden Krankheit gleichgesetzt. Wenn dies nicht der Wahrheit entspricht, dann wurde die entsprechende Variable zwar misklassifiziert, sie hat jedoch einen Wert. Die oben genannten Validierungsstudien quantifizieren das Ausmaß der Missklassifikation und ermöglichen eine Einschätzung, ob eine Studie noch valide durchgeführt und die Effektschätzung

durch die Modellierung des potenziellen Bias korrigiert werden kann [19, 29].

In elektronischen Patientenakten werden u. a. Testwerte erfasst, die nicht einfach auf einen festgelegten Wert gesetzt werden können. Die Tatsache, dass ein bestimmter Test vom behandelnden Arzt nicht angeordnet wurde, ist an sich bereits informativ. Die daraus resultierende Strategie, einen Indikator für fehlende Variablenwerte in die Analyse mit aufzunehmen, hat sich, mit Ausnahme von Extremfällen, als nützlich erwiesen [42, 43]. Andere Strategien hierfür sind die Imputation des Mittelwerts aller beobachteten Werte oder das Ersetzen des fehlenden Werts mit dem zuletzt beobachteten Wert, die beide nicht zufriedenstellend sind, aber häufig angewendet werden [41]. Multiple Imputation, die eine Reihe beobachteter Variablen hinzuzieht, nimmt an, dass letztere mit dem fehlenden Wert assoziiert sind, eine Annahme, die häufig haltlos und empirisch nicht überprüfbar ist.

Datenanalyse

In longitudinalen Studien können eine Reihe von kausalen Parametern geschätzt werden. Im vorliegenden Beitrag werden die für RWE-Studien am relevantesten Parameter berücksichtigt. Bei der Wahl des analytischen Ansatzes muss zwangsweise ein Kompromiss zwischen der klinischen Relevanz verschiedener Zielgrößen und der Durchführbarkeit einer unverzerrten Schätzung der gewählten Zielgröße gefunden werden.

Interessierender kausaler Effekt

Der „As-treated"-Effekt

Beim „As-treated"-Effekt handelt es sich um den Behandlungseffekt einer im Rahmen der Studie zu untersuchenden Behandlung, die initiiert und weitergeführt wird. Die Beobachtungszeit der Patienten wird nach einem Abbruch der initialen Behandlung zensiert. Der numerische Wert des „As-treated"-Effekts aus einer Studie berücksichtigt dementsprechend die Dauer der Behandlungspersistenz. In den meisten Fällen ist der „As-treated"-Effekt sowohl für Patienten als auch Ärzte von großem Interesse, da er Aufschluss über den zu erwartenden Behandlungseffekt gibt, während sich der Patient tatsächlich in Behandlung befindet.

Der Effekt komplexer Behandlungsstrategien

Bei vielen chronischen Erkrankungen wird empfohlen, eine Arzneimittelbehandlung in Abhängigkeit von klinischen Merkmalen zu beginnen, zu beenden, zu wechseln oder die Dosierung zu ändern. Daher kann es im Interesse stehen, anstelle des Behandlungseffekts mit einem einzelnen Medikament, den Effekt einer longitudinalen Behandlungsstrategie zu analysieren.

Der „As-started"-Effekt

Beim „As-started"-Effekt handelt es sich um den Behandlungseffekt der initialen Behandlung, unabhängig davon, ob diese über einen bestimmten Zeitraum fortgesetzt wurde. Dies entspricht dem „Intention-to-treat"-Prinzip in RCT. Das Ausmaß des „As-started"-Effekts im Nachbeobachtungszeitraum einer pharmakoepidemiologischen Studie hängt von dem spezifischen Grad der Abweichungen der initialen Behandlung ab. Wenn Patienten eine Behandlung abbrechen, wird ihr Expositionsstatus weiterhin entsprechend der initialen Behandlungswahl kategorisiert. Diese Vorgehensweise vermeidet Schwierigkeiten, die sich aus der informativen Zensierung ergeben (d. h. Teilnehmer sind aus Gründen, die mit dem Studienergebnis zusammenhängen zensiert worden), führt aber zu einer Expositionsmissklassifikation über die Summe der Personenzeit.

Wahl der Behandlung und Confounding kausaler Behandlungseffekte

Ärzte treffen eine Behandlungswahl unter Berücksichtigung des Erkrankungsschweregrades und der zum Zeitpunkt der Verordnung verfügbaren prognostischen Informationen aus. Die Faktoren, die diese Wahl beeinflussen, variieren von Arzt zu Arzt sowie im Laufe der Zeit und umfassen häufig klinische, funktionelle oder verhaltensbezogene Patientenmerkmale, die in Datenbanken des Gesundheitswesens möglicherweise nicht vollständig erfasst werden. Wenn solche prognostischen Faktoren zwischen den behandelten Patienten und den Vergleichspatienten nicht ausgeglichen sind, kann eine fehlende statistische Kontrolle dieser Faktoren zu Verzerrungen führen. Da die Wahl der Behandlung je nach Schweregrad und Prognose der Erkrankung ein integraler Bestandteil der medizinischen Praxis ist, kann die daraus resultierende Verzerrung (Bias) sehr stark sein. Das Confounding, das durch eine selektive Behandlungswahl in der Praxis entsteht, wird manchmal spezifischer als „confounding by indication", „confounding by contraindication", „channeling bias" oder „healthy user bias" bezeichnet. All diese Bezeichnungen zielen auf dasselbe zugrundeliegende Problem ab.

Die Analyse vergleichbarer Patienten

Beschränkung auf ähnliche Patienten

Eine Vergleichbarkeit der Behandlungsgruppen in der Abwesenheit von randomisierter Behandlungseinteilung herzustellen ist ein wichtiges Ziel, das verschiedentlich angegangen werden kann. Beschränkung ist ein gängiges und wirksames Analysetool, um Behandlungsgruppen vergleichbarer zu machen, damit unbekanntes bzw. nicht gemessenes Confounding (residuales Confounding) minimiert werden kann. Einige dieser Beschränkungen liegen auf der Hand, da sie anhand expliziter Kriterien festgelegt werden: Beispielsweise die Beschränkung der Studienpopulation auf Patienten, die 65 Jahre oder älter sind und an Demenz leiden zur Untersuchung der Sicherheit von antipsychotischen Medikamenten, die zur Behandlung von Verhaltensstörungen in eben dieser Bevölkerungsgruppe eingesetzt werden. Andere Beschränkungen, wie ein Matching hinsichtlich eines zusammenfassenden Confounder-Scores (entweder eines Propensity Scores oder eines Risikoscores), werden in der Pharmakoepidemiologie häufig verwendet. Es ist wichtig, die spezifischen Gründe für die entsprechenden Beschränkungen zu verstehen, damit die Verringerung von Confounding gegen

die Einschränkung der Generalisierbarkeit der Ergebnisse abgewogen werden kann [34].

Propensity-Score-Analysen

Propensity Scores (PS) können eine große Anzahl von Kovariaten effizient adjustieren, selbst wenn das zu untersuchende Outcome selten auftritt. Daher haben sich PS-Analysen als praktisches und effektives Mittel zur Adjustierung einer großen Anzahl potenzieller Confounder in Wirksamkeitsstudien basierend auf Real-World-Daten erwiesen. Sie passen zum Paradigma des „target trials" (auf welchen bereits hingewiesen wurde), da der PS den Randomisierungsprozess auf Grundlage von Beobachtungsdaten nachbildet (emuliert). In einem „New-user"-Kohortendesign ist ein PS die geschätzte Wahrscheinlichkeit, mit Behandlung A anstelle von Behandlung B zu beginnen, in Abhängigkeit von allen Patientenmerkmalen, die vor Beginn der Behandlung beobachtet wurden. PS lassen sich mittels logistischer Regression unkompliziert schätzen und Strategien für die Auswahl der dabei einzubeziehenden Variablen wurden an anderer Stelle beschrieben [2]. Sobald ein PS auf Grundlage beobachteter Kovariaten geschätzt wurde, gibt es mehrere Möglichkeiten, ihn in einem zweiten Schritt zur Reduzierung von Confounding anzuwenden. Typische Strategien schließen die Adjustierung für Quintile oder Dezile des Scores mit oder ohne Trimming, Matching, Feinstratifizierung oder Gewichtung mit dem PS ein [17].

Im Rahmen von Kohortenstudien bietet das PS-Matching mehrere Vorteile, die den vermeintlichen Nachteil aufwiegen können, dass manchmal nicht der gesamte Datensatz genutzt wird, weil nicht alle in Frage kommenden Patienten einen vergleichbaren Match finden konnten. Das Matching schließt Patienten in den extremen PS-Bereichen aus, in denen die Behandlungswahl deutlich eingeschränkt ist, z. B. alle Patient werden mit Therapie A behandelt. Wenn solche Patienten aus der Analyse ausgeschlossen werden, wird das residuelle Confounding verringert und der Fokus auf Patienten, denen eine Behandlungswahl offen steht, erhöht die klinische Relevanz [45]. Im Gegensatz zu traditionellen Outcome-Modellen ermöglichen PS-gematchte Analysen, insbesondere das „fixed ratio matching", dem Forschenden, die in der Studienpopulation erreichte Balanciertheit der Kovariaten nachzuweisen. Post-matching-C-Statistiken oder standardisierte Differenzen der Kovariaten haben in PS-gematchten Analysen an Popularität gewonnen [7]. In Kohortenstudien erfordert das „fixed ratio matching", wie das häufig angewandte 1:1-PS-Matching, vereinfachte Analysen, um ein unverzerrtes Ergebnis zu erzielen. In Settings mit sehr wenigen Ereignissen kann eine feinere Stratifizierung bevorzugt werden [5].

Jegliche vor der Exposition erfassten Patientenmerkmale können als potenzielle Confounder betrachtet werden. Sofern im Falle der Verwendung von Sekundärdaten eine optimale Erfassung entsprechender Merkmale nicht möglich ist, kann durch die Erfassung und Adjustierung anhand beobachtbarer Proxies das Confounding reduziert werden. Der unbeobachtbare Confounder wird zu dem Maße adjustiert, in dem ein Proxy mit dem entsprechenden Confounder korreliert [4, 10]. Beispiele für angemessene Proxies sind die Nutzung von Sauerstoffflaschen (korreliert mit körperlicher Gebrechlichkeit), die regelmäßige Inanspruchnahme von Präventionsmaßnahmen (korreliert mit gesundheitsbewusstem Verhalten) oder die Einnahme von Medikamenten zur Senkung des Blutzuckerspiegels (korreliert mit Hb_{A1c}-Messwerten) usw. Hierdurch kann ein hochdimensionalen Kovariatenraum mit mehreren tausend Kovariaten gebildet werden, von denen einige echte Confounder sind [35]. Techniken zur Verringerung der Anzahl an Variablen reduzieren die Anzahl an Kovariaten (die möglicherweise Confounder sind) von mehreren tausend auf einige hundert Kovariaten (die mit hoher Wahrscheinlichkeit tatsächlich Confounder sind), bevor sie in das PS-Modell eingehen [18, 35, 36]. Der sich daraus resultierende hochdimensionale PS ist im Hinblick auf die Verringerung von Verzerrungen bei einer Reihe von Forschungsfragen häufig überlegen [31, 52]. Obwohl selbst eine hochdimensionale Adjustierung von Confoundern nicht mit einer Randomisierung vergleichbar ist, kann in vielen Fällen demonstriert werden, dass Kausalzusammenhänge mit Real-World-Daten identifiziert werden können.

Subgruppenanalysen und Behandlungseffektmodifikation

Große Real-World-Datenquellen erlauben, Analysen nach vielen Faktoren zu stratifizieren, die für die verschreibenden Ärzte und ihre Patienten relevant sind. Allgemeine Empfehlungen zur Untersuchung heterogener Behandlungseffekte gelten für RWE- ebenso wie für RCT-Studien [20]. Besonders problematisch bleiben Post-hoc-Tests auf Effektmodifikation, die trotz recht konservativer statistischer Tests auf Interaktion zu falsch-positiven Ergebnissen führen können. Anzeichen für eine Effektmodifikation sollten in Folgestudien basierend auf anderen Datenquellen bestätigt werden [51].

Zusammenfassung

Dieser Artikel hat das Handwerkszeug und die generellen Angehensweisen von RWE-Studien beschrieben, um eine Forschungsfrage mit kausaler Interpretation beantworten zu können. Dem zugrunde liegt ein Verständnis der biologischen Natur und der medizinischen Versorgungspraxis, die in einem abstrakten Model, dem Studiendesign und den notwendigen Messungen abgebildet und schließlich in einer statistischen Analyse zusammengefasst werden. Dies geht mit vielen Annahmen und Vereinfachungen einer komplexen Welt einher. Produzierer von Evidenz in der Medizin, experimentell genauso wie Real World und Entscheidungsträger, die gemeinsam das medizinische Versorgungsunternehmen verbessern wollen, werden sich dieser Ungewissheit bewusst sein und entsprechend abgewogen kommunizieren und handeln.

Fazit für die Praxis

- **Die Durchführung von RWE(Real World Evidence)-Studie sollte stets das Ziel verfolgen, zu kausalen Schlussfolgerungen in Bezug auf**

- **die Wirksamkeit der untersuchten Behandlung zu gelangen.**
- **RWE-Studien untersuchen die Wirksamkeit von medizinischen Produkten oder Interventionen im klinischen Versorgungsalltag und stellen eine Ergänzung zur Evidenz von randomisierten Studien dar, die deren Wirksamkeit im kontrollierten Forschungssetting untersuchen.**
- **Der „Target trial"-Ansatz schafft Klarheit bei der Planung und Interpretation von RWE-Studien. In Verbindung mit modernen Methoden der Epidemiologie und Biostatistik hilft dieser Ansatz außerdem, vom Untersucher verursachte Verzerrungen zu vermeiden und die Übereinstimmung von Studiendesign und Forschungsfrage sicherzustellen.**
- **Die Arbeit mit Sekundärdaten aus dem Gesundheitswesen stellt eine Herausforderung hinsichtlich der Datenqualität und -vollständigkeit dar, die bei der Design- und Analysestrategie von RWE-Studien berücksichtigt werden muss.**
- **Eine enge Zusammenarbeit zwischen klinischen Experten, Experten der verwendeten Datenquelle und Epidemiologen und Biostatistikern ist für den Erfolg entscheidend.**

Korrespondenzadresse

Professor Dr. Dr. Sebastian Schneeweiss
Abteilung für Pharmakoepidemiologie und Pharmakoökonomie, Klinik der Inneren Medizin, Brigham und Women's Hospital, Harvard Medical School
1 Brigham Circle (Suite 3030), 02120 Boston, MA, USA
schneeweiss@post.harvard.edu

Einhaltung ethischer Richtlinien

Interessenkonflikt. S. Schneeweiss ist an Forschungsförderungen für das Brigham and Women's Hospital von UCB und Boehringer Ingelheim beteiligt, die nicht im Zusammenhang mit dem Thema dieser Studie stehen. Er ist Berater und Anteilseigner von Aetion Inc. einem Softwarehersteller, von dem er auch Anteile besitzt. Diese Interessen wurden vom Brigham and Women's Hospital offengelegt, geprüft und genehmigt und sind in Übereinstimmung mit den institutionellen Compliance-Richtlinien.

Für diesen Beitrag wurden von den Autor/-innen keine Studien an Menschen oder Tieren durchgeführt. Für die aufgeführten Studien gelten die jeweils dort angegebenen ethischen Richtlinien.

Literatur

1. Brookhart MA, Wang PS, Solomon DH et al (2006) Evaluating short-term drug effects using a physician-specific prescribing preference as an instrumental variable. Epidemiology 17(3):268–275. https://doi.org/10.1097/01.ede.0000193606.58671.c5
2. Brookhart MA, Schneeweiss S, Rothman KJ et al (2006) Variable selection for propensity score models. Am J Epidemiol 163(12):1149–1156. https://doi.org/10.1093/AJE/KWJ149
3. Chan KA, Andrade SE, Boles M et al (2000) Inhibitors of hydroxymethylglutaryl-coenzyme A reductase and risk of fracture among older women. Lancet 355(9222):2185–2188. https://doi.org/10.1016/S0140-6736(00)02400-4
4. Christopeit N (2003) Wooldridge, J. M.: econometric analysis of cross section and panel data. XXIII, 752 pp. MIT press, Cambridge, Mass., 2002. Hardcover £ 37.50. J Econ 80(2):206–209. https://doi.org/10.1007/s00712-003-0589-6
5. Desai RJ, Rothman KJ, Bateman BT et al (2017) A propensity-score-based fine stratification approach for confounding adjustment when exposure is infrequent. Epidemiology 28(2):249–257. https://doi.org/10.1097/EDE.0000000000000595
6. Faillie J-L, Yu OH, Yin H et al (2016) Association of bile duct and gallbladder diseases with the use of Incretin-based drugs in patients with type 2 diabetes mellitus. JAMA Intern Med 176(10):1474–1481. https://doi.org/10.1001/jamainternmed.2016.1531
7. Franklin JM, Rassen JA, Ackermann D et al (2014) Metrics for covariate balance in cohort studies of causal effects. Stat Med 33(10):1685–1699. https://doi.org/10.1002/sim.6058
8. Franklin JM, Glynn RJ, Martin D et al (2019) Evaluating the use of nonrandomized real-world data analyses for regulatory decision making. Clin Pharmacol Ther 105(4):867–877. https://doi.org/10.1002/cpt.1351
9. Funk MJ, Landi SN (2014) Misclassification in administrative claims data: quantifying the impact on treatment effect estimates. Curr Epidemiol Rep 1(4):175–185. https://doi.org/10.1007/s40471-014-0027-z
10. Gelman A, Carlin JB, Stern HS, Rubin DB (1995) Bayesian data analysis. Chapman and Hall/CRC
11. Gokhale M, Stürmer T, Buse JB (2020) Real-world evidence: the devil is in the detail. Diabetologia 63(9):1694–1705. https://doi.org/10.1007/s00125-020-05217-1
12. Grodstein F, Stampfer MJ, Manson JE et al (1996) Postmenopausal estrogen and progestin use and the risk of cardiovascular disease. N Engl J Med 335(7):453–461. https://doi.org/10.1056/NEJM199608153350701
13. Heart Protection Study Collaborative Group (2002) MRC/BHF heart protection study of cholesterol lowering with simvastatin in 20,536 high-risk individuals: a randomised placebo-controlled trial. Lancet 360(9326):7–22. https://doi.org/10.1016/S0140-6736(02)09327-3
14. Hernán MA, Robins JM (2016) Using big data to emulate a target trial when a randomized trial is not available. Am J Epidemiol 183(8):758–764. https://doi.org/10.1093/aje/kwv254
15. Hernán MA, Alonso A, Logan R et al (2008) Observational studies analyzed like randomized experiments: an application to postmenopausal hormone therapy and coronary heart disease. Epidemiology 19(6):766–779. https://doi.org/10.1097/EDE.0b013e3181875e61
16. Hernán MA, Sauer BC, Hernández-Díaz S et al (2016) Specifying a target trial prevents immortal time bias and other self-inflicted injuries in observational analyses. J Clin Epidemiol 79:70–75. https://doi.org/10.1016/j.jclinepi.2016.04.014
17. Johnson ES, Bartman BA, Briesacher BA et al (2013) The incident user design in comparative effectiveness research. Pharmacoepidemiol Drug Saf 22(1):1–6. https://doi.org/10.1002/pds.3334
18. Karim ME, Pang M, Platt RW (2018) Can we train machine learning methods to outperform the high-dimensional propensity score algorithm? Epidemiology 29(2):191–198. https://doi.org/10.1097/EDE.0000000000000787
19. Lash TL, Fox MP, MacLehose RF et al (2014) Good practices for quantitative bias analysis. Int J Epidemiol 43(6):1969–1985. https://doi.org/10.1093/ije/dyu149
20. Lesko CR, Henderson NC, Varadhan R (2018) Considerations when assessing heterogeneity of treatment effect in patient-centered outcomes research. J Clin Epidemiol 100:22–31. https://doi.org/10.1016/j.jclinepi.2018.04.005
21. Maclure M (2007) 'Why me?' versus 'why now?'—differences between operational hypotheses in case-control versus case-crossover studies. Pharmacoepidemiol Drug Saf 16(8):850–853. https://doi.org/10.1002/pds.1438
22. Malone DC, Brown M, Hurwitz JT et al (2018) Real-world evidence: useful in the real world of US payer decision making? How? When? And what studies? Value Health 21(3):326–333. https://doi.org/10.1016/j.jval.2017.08.3013
23. Moran LV, Ongur D, Hsu J et al (2019) Psychosis with methylphenidate or amphetamine in patients with ADHD. N Engl J Med 380(12):1128–1138. https://doi.org/10.1056/NEJMoa1813751
24. Patorno E, Goldfine AB, Schneeweiss S et al (2018) Cardiovascular outcomes associated with canagliflozin versus other non-gliflozin antidiabetic drugs: population based cohort study. BMJ 360:k119. https://doi.org/10.1136/bmj.k119

25. Patorno E, Schneeweiss S, Gopalakrishnan C et al (2019) Using real-world data to predict findings of an ongoing phase IV cardiovascular outcome trial: cardiovascular safety of linagliptin versus glimepiride. Diabetes Care 42(12):2204–2210. https://doi.org/10.2337/dc19-0069
26. Ray WA (2003) Evaluating medication effects outside of clinical trials: new-user designs. Am J Epidemiol 158(9):915–920. https://doi.org/10.1093/aje/kwg231
27. Rimm EB, Stampfer MJ, Ascherio A et al (1993) Vitamin E consumption and the risk of coronary heart disease in men. N Engl J Med 328(20):1450–1456. https://doi.org/10.1056/NEJM199305203282004
28. Rothman KJ, Poole C (1988) A strengthening programme for weak associations. Int J Epidemiol 17(4):955–959. https://doi.org/10.1093/ije/17.4.955
29. Schneeweiss S (2006) Sensitivity analysis and external adjustment for unmeasured confounders in epidemiologic database studies of therapeutics. Pharmacoepidemiol Drug Saf 15(5):291–303. https://doi.org/10.1002/pds.1200
30. Schneeweiss S (2010) A basic study design for expedited safety signal evaluation based on electronic healthcare data. Pharmacoepidemiol Drug Saf 19(8):858–868. https://doi.org/10.1002/pds.1926
31. Schneeweiss S (2018) Automated data-adaptive analytics for electronic healthcare data to study causal treatment effects. Clin Epidemiol 10:771–788. https://doi.org/10.2147/CLEP.S166545
32. Schneeweiss S (2019) Real-world evidence of treatment effects: the useful and the misleading. Clin Pharmacol Ther 106(1):43–44. https://doi.org/10.1002/cpt.1405
33. Schneeweiss S, Avorn J (2005) A review of uses of health care utilization databases for epidemiologic research on therapeutics. J Clin Epidemiol 58(4):323–337. https://doi.org/10.1016/j.jclinepi.2004.10.012
34. Schneeweiss S, Patrick AR, Stürmer T et al (2007) Increasing levels of restriction in pharmacoepidemiologic database studies of elderly and comparison with randomized trial results. Med Care 45(10):S131–42. https://doi.org/10.1097/MLR.0b013e318070c08e
35. Schneeweiss S, Rassen JA, Glynn RJ et al (2009) High-dimensional propensity score adjustment in studies of treatment effects using health care claims data. Epidemiology 20(4):512–522. https://doi.org/10.1097/EDE.0b013e3181a663cc
36. Schneeweiss S, Eddings W, Glynn RJ et al (2017) Variable selection for confounding adjustment in high-dimensional covariate spaces when analyzing healthcare databases. Epidemiology 28(2):237–248. https://doi.org/10.1097/EDE.0000000000000581
37. Setoguchi S, Warner Stevenson L, Stewart GC et al (2014) Influence of healthy candidate bias in assessing clinical effectiveness for implantable cardioverter-defibrillators: cohort study of older patients with heart failure. BMJ 348:g2866. https://doi.org/10.1136/bmj.g2866
38. Shadish WR, Clark MH, Steiner PM (2008) Can nonrandomized experiments yield accurate answers? A randomized experiment comparing random and nonrandom assignments. J Am Stat Assoc 103(484):1334–1344. https://doi.org/10.1198/016214508000000733
39. Smeeth L, Douglas I, Hall AJ et al (2009) Effect of statins on a wide range of health outcomes: a cohort study validated by comparison with randomized trials. Br J Clin Pharmacol 67(1):99–109. https://doi.org/10.1111/j.1365-2125.2008.03308.x
40. Stampfer MJ, Hennekens CH, Manson JE et al (1993) Vitamin E consumption and the risk of coronary disease in women. N Engl J Med 328(20):1444–1449. https://doi.org/10.1056/NEJM199305203282003
41. Sterne JAC, White IR, Carlin JB et al (2009) Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. BMJ 338:b2393. https://doi.org/10.1136/bmj.b2393
42. Vach W, Blettner M (1991) Biased estimation of the odds ratio in case-control studies due to the use of ad hoc methods of correcting for missing values for confounding variables. Am J Epidemiol 134(8):895–907. https://doi.org/10.1093/oxfordjournals.aje.a116164
43. Vach W, Blettner M (1995) Logistic regression with incompletely observed categorical covariates—investigating the sensitivity against violation of the missing at random assumption. Stat Med 14(12):1315–1329. https://doi.org/10.1002/sim.4780141205
44. VanderWeele TJ (2019) Principles of confounder selection. Eur J Epidemiol 34(3):211–219. https://doi.org/10.1007/s10654-019-00494-6
45. Walker A, Patrick, Lauer M et al (2013) A tool for assessing the feasibility of comparative effectiveness research. CER. https://doi.org/10.2147/CER.S40357
46. Webster-Clark M, Stürmer T, Wang T et al (2021) Using propensity scores to estimate effects of treatment initiation decisions: state of the science. Stat Med 40(7):1718–1735. https://doi.org/10.1002/sim.8866
47. West SL, Savitz DA, Koch G et al (1995) Recall accuracy for prescription medications: self-report compared with database information. Am J Epidemiol 142(10):1103–1112. https://doi.org/10.1093/oxfordjournals.aje.a117563
48. Wilchesky M, Tamblyn RM, Huang A (2004) Validation of diagnostic codes within medical services claims. J Clin Epidemiol 57(2):131–141. https://doi.org/10.1016/S0895-4356(03)00246-4
49. Winkelmayer WC, Setoguchi S, Levin R et al (2008) Comparison of cardiovascular outcomes in elderly patients with diabetes who initiated rosiglitazone vs pioglitazone therapy. Arch Intern Med 168(21):2368–2375. https://doi.org/10.1001/archinte.168.21.2368
50. Yusuf S, Dagenais G, Pogue J et al (2000) Vitamin E supplementation and cardiovascular events in high-risk patients. N Engl J Med 342(3):154–160. https://doi.org/10.1056/NEJM200001203420302
51. Zhang Y, Laber EB, Tsiatis A et al (2015) Using decision lists to construct interpretable and parsimonious treatment regimes. Biometrics 71(4):895–904. https://doi.org/10.1111/biom.12354
52. Zhou M, Wang SV, Leonard CE et al (2017) Sentinel modular program for propensity score-matched cohort analyses: application to glyburide, glipizide, and serious hypoglycemia. Epidemiology 28(6):838–846. https://doi.org/10.1097/EDE.0000000000000709