

Vom Datenchaos zum strukturierten Modell – Wie sich Daten zähmen lassen

CHRISTEL WEISS

Medizinische Fakultät Mannheim der Universität Heidelberg

Zusammenfassung

Dieser Beitrag stellt sich der Herausforderung, die beiden Begriffe „Chaos“ und „Struktur“ zu einem harmonischen Ganzen zu verbinden. Es wird dargelegt, wie aus dem Gedankenchaos im Kopf eines Forschers ein klares Konzept entsteht, wie Chaos bereits bei der Studienplanung vermieden werden kann und wie es mittels statistischer Methoden gelingt, einem unstrukturierten Datenchaos ein strukturiertes Profil zu verleihen. Das Entdecken dieser Strukturen ermöglicht es, Zusammenhänge zu erkennen und neues Wissen zu erlangen. Diese Vorgehensweise wird anhand ausgewählter Therapiestudien aus zwei Jahrhunderten erläutert.

1 Einleitung

Chaos und Struktur: Diese beiden Begriffe wecken höchst unterschiedliche Assoziationen. Das Chaos ist nach der griechischen Mythologie der Urstoff, aus dem die Welt hervorgegangen ist. Ursprünglich bedeutet das altgriechische Wort *Χαος* „gähnende Leere“. Nach alttestamentarischen Überlieferungen erschuf Gott aus dem Chaos die Welt. Im übertragenen Sinne ist der Begriff „Chaos“ negativ konnotiert: Er steht für Unordnung, Ineffizienz oder eine verfahrenere Situation.

Dagegen symbolisiert der Begriff „Struktur“ Eigenschaften wie Ordnung, Klarheit oder Übersichtlichkeit. Er leitet sich ab vom lateinischen „structura“ (Bauart,

Aufbau) und beschreibt die Art, wie die Teile eines komplexen Gefüges miteinander in Beziehung stehen. In zahlreichen Naturwissenschaften ist „Struktur“ zu einem Grundbegriff avanciert: Man denke etwa an Zellstrukturen oder an chemische Strukturen, die den Aufbau einer Zelle, eines Atoms oder Moleküls beschreiben, oder an mathematische Strukturen wie beispielsweise eine Boolesche Algebra, die die Verknüpfungen innerhalb einer Menge definiert. Auch in den Sozial- und Wirtschaftswissenschaften ist „Struktur“ ein Begriff von zentraler Bedeutung. Eine Gemeinschaft von menschlichen Individuen ist durch Sozialstrukturen gekennzeichnet. Ein Staat wird charakterisiert durch seine Bevölkerungsstruktur sowie seine ökonomische, Agrar- und Infrastruktur. In kleineren Sozialgebilden wie Familien, Kollegien und Vereinen regeln Strukturen die Beziehungen der Mitglieder untereinander, deren Funktionen und Verantwortung innerhalb der Gemeinschaft. Auch im Leben eines Individuums sind Strukturen unerlässlich: Niemand ist gerne über einen längeren Zeitraum dem totalen Chaos ausgeliefert. Deshalb ist jeder Mensch bemüht, dem Chaos ein gewisses Maß an Ordnung abzuringen – ansonsten würde er daran zugrunde gehen. Strukturen gewähren Orientierung, sie bieten Sicherheit und Stabilität.

Chaos per se ist ein unbefriedigender Zustand. Andererseits ist er die Voraussetzung dafür, dass eine Struktur entstehen kann [6]. Ein schlichtes Nebeneinander von Elementarteilchen eines Atoms, Organismen einer Zelle oder einzelnen Personen in einem sozialen Verbund scheint nicht sinnvoll zu sein. Erst die Offenlegung der Struktur, also der Art, wie Objekte oder Menschen miteinander agieren, zeigt die Komplexität der Beziehungsgeflechte auf und ermöglicht tiefer gehende Erkenntnisse.

Wie lässt sich eine Struktur im Chaos aufspüren? Um Ordnungsstrukturen zu erkennen, müssen Regelmäßigkeiten gefunden werden. Dies gilt für die hohe Wissenschaft ebenso wie für die Niederungen des Alltags. Die Entwirrung eines Chaos beginnt mit der Beschreibung der einzelnen Elemente. Durch die Entwicklung der Elektronenmikroskopie wurde es möglich, die Bestandteile einer Zelle (Membran, Plasma, Organellen u. a.) zu erkennen und zu benennen. Zu Beginn des 20. Jahrhunderts entdeckten Wissenschaftler wie Ernest Rutherford (1871–1937) oder Niels Bohr (1885–1962), dass ein Atom aus einem Kern und Elektronen besteht. Soziale Wesen lassen sich durch Merkmale wie Alter, Bildung oder sozio-ökonomischer Status beschreiben. Das Chaos auf einem Schreibtisch ist nichts

weiter als eine Ansammlung von Büchern, Zetteln, Büroutensilien, wichtigen Dokumenten, unerledigten Aufgaben etc.

Freilich erklären einfache Beschreibungen nicht die Funktionsweise der einzelnen Elemente und deren Interaktionen untereinander. Gleichwohl ist dies ein erster und wichtiger Schritt, um Licht ins Chaos zu bringen. Die Beschreibung regt die Kreativität an und verleitet dazu, zu ordnen und zu sortieren, Fragen zu stellen, zu handeln und zu forschen. Welchem Zweck dienen die Gegenstände auf einem Schreibtisch? Welche Beziehungen verbinden die Mitglieder einer Gesellschaft untereinander? Wie sind die Elemente in einem Atom oder einer Zelle angeordnet? Welche Funktion haben sie? In nachfolgenden Schritten wird man sich bemühen, Antworten zu finden, Strukturen im Chaos zu entdecken und darin verborgene Geheimnisse zu ergründen.

Was haben diese Überlegungen mit Studien, Daten und statistischen Analysen zu tun? In empirischen Wissenschaften wie beispielsweise der Medizin oder der Psychologie ist die Statistik als wissenschaftsübergreifende Disziplin unerlässlich, um Studien zu planen und durchzuführen, Daten zu analysieren und die Ergebnisse zu interpretieren. Chaos macht sich zuweilen breit: Noch vor Studienbeginn im Kopf eines Wissenschaftlers (etwa eines Kliniklers), der seine Gedanken ordnen muss, um eine konkrete Hypothese zu formulieren und eine Studie planen zu können. Später wird der Statistiker mit einem Chaos konfrontiert sein, denn Daten im Urzustand stellen im wahrsten Sinne des Wortes ein Chaos dar. Das Studienziel besteht im Wesentlichen darin, das Chaos aufzulösen, um eine Antwort auf die der Studie zugrundeliegende Fragestellung zu finden und so zu neuen Erkenntnissen zu gelangen.

2 Klinische Studien: Der lange Weg des therapeutischen Fortschritts

2.1 Vom Gedankenchaos zum klaren Konzept

Bei einer empirischen Studie sind in aller Regel Wissenschaftler aus mehreren Fachrichtungen beteiligt. Neben einem Vertreter aus dem Wissenschaftsgebiet, in dem die Studie durchgeführt wird (etwa eines Kliniklers), bedarf es eines Statistikers, der bei der Studienplanung, der Datenanalyse sowie der Interpretation der Ergebnisse behilflich ist. Dies sei am Beispiel einer Therapiestudie erläutert (wenngleich andere Studientypen wie Risiko-, Diagnose- oder Präventionsstudien grundsätzlich ähnliche Überlegungen erfordern).

Was steht am Beginn einer solchen Studie? Ein Forscher entwickelt eine Idee für ein neues therapeutisches Konzept – sei es aufgrund der Kenntnis der Krankheitsmechanismen oder einer zufälligen Entdeckung, aufgrund einer spontanen Idee oder nach langjährigen Erfahrungen. Er möchte diese neue Therapie überprüfen und – falls sie sich als tauglich erweisen sollte – dafür Sorge tragen, dass sie später bei der Patientenversorgung Anwendung findet. Ein neues Arzneimittel, ein neu entwickeltes chirurgisches Verfahren oder die Optimierung einer bereits zugelassenen Therapie soll letztlich den Patienten zugutekommen. Dieser Nutzen ist im Rahmen einer Studie nachzuweisen. Doch wie soll der Nutzen gemessen werden? Eine Antwort auf diese Frage ist keineswegs trivial. Es ist anzunehmen, dass dem Initiator einer Studie ein Wirrwarr von Ideen durch den Kopf huscht, ehe er einen konkreten Plan erarbeitet. Tatsächlich sind die Wirkungen einer Therapie, die Vorgänge und Wechselwirkungen im menschlichen Körper so komplex, dass sie sich einer genauen Beschreibung entziehen. Deshalb fällt es zunächst schwer, eine präzise Fragestellung zu formulieren. Chaos ist also quasi vorprogrammiert. Um dieses Gedankenchaos aufzulösen, sollte der Forscher strukturiert vorgehen und folgende Fragen klären:

- **Gegenstand der Forschung:** Zunächst ist die Forschungsthematik inhaltlich einzugrenzen. Handelt es sich bei der zu evaluierenden Therapie um ein Medikament, einen chirurgischen Eingriff, ein Medizingerät, eine Psychotherapie oder eine andere Therapieart? Wird ein neues Produkt getestet oder ist die zu prüfende Therapie bereits auf dem Markt?
- **Stand der Forschung:** Der Studienverantwortliche sollte rekapitulieren, welche Voruntersuchungen bezüglich des Forschungsvorhabens durchgeführt wurden und welche Erkenntnisse vorliegen. Sodann ist zu klären: Welche Phase des Entwicklungsprozesses steht an? Zudem sollte mittels einer umfassenden Literaturrecherche der aktuelle Stand der Wissenschaft bezüglich dieser Thematik eruiert werden.
- **Studienziel:** Welches Ziel steht im Vordergrund? Die Beantwortung dieser Frage hängt wesentlich von der Entwicklungsphase der Therapie ab. Bei der Entwicklung eines Medikaments müssen diverse Phasen durchlaufen werden, ehe die Zulassung beantragt werden kann [3]. In der vorklinischen Phase wird das Arzneimittel mittels Labor- oder Tierexperimenten bezüglich Wirkmechanismus und Toxikologie untersucht. In Phase I wird es an

gesunden Probanden getestet, um Informationen bezüglich Pharmakodynamik (Wirkung des Arzneimittels im Organismus) und Pharmakokinetik (Verteilung und Verstoffwechslung im Körper) zu gewinnen. In Phase II wird das Medikament erstmals Patienten verabreicht, um die therapeutische Wirksamkeit zu prüfen und unerwünschte Nebenwirkungen zu erfassen. In Phase III wird das neue Medikament mit einer anderen Therapie verglichen, um das Wirkungsvermögen unter Studienbedingungen zu prüfen („Efficacy“). Wenn all diese Phasen erfolgreich abgeschlossen sind, kann die Zulassung beantragt werden. Danach beginnt die Phase IV, in der die Compliance („Einnahme-Treue“) unter Alltagsbedingungen („Effectiveness“), seltene Nebenwirkungen und die Bewährung des Medikaments bei speziellen Patientengruppen erforscht sowie Untersuchungen zu Therapieoptimierung durchgeführt werden. – Soweit in aller Kürze der Prozess bei der Entwicklung eines Arzneimittels, der im Schnitt insgesamt – von der Idee bis zur Marktreife – mehr als zehn Jahre andauert. Bei anderen Therapieformen wie beispielsweise Medizinprodukten (dazu zählen Implantate, Instrumente, Diagnostika) sind die Vorgaben weniger streng. Aber auch sie sind klinisch zu bewerten, ehe sie routinemäßig angewandt werden können.

- **Zielgrößen:** Die Wahl relevanter Zielgrößen ergibt sich aus dem Studienziel und hängt wie dieses von der Entwicklungsphase des Produkts und der zu therapierenden Erkrankung ab. Studienärzte sprechen von „Endpunkten“ oder verwenden den englischen Begriff „Outcomes“. In der vorklinischen Phase und in der Phase I werden Endpunkte verwendet, die sich mit molekularen oder physiologischen Termini beschreiben lassen. In der Phase II werden häufig Surrogatendpunkte erhoben, um die Wirkung der Therapie zu messen [2]. Diese Ersatzgrößen sind hilfreich, da sie meist schon nach kurzer Zeit gemessen werden können, während sich der eigentlich angestrebte Nutzen einer Behandlung oft erst nach Jahren oder Jahrzehnten erweist. So stellen zum Beispiel Biomarker (z. B. Blutdruck- oder Cholesterinwerte) Surrogatparameter dar bei Therapieoptionen, die auf eine Verlängerung der Lebenszeit oder Vermeidung von Herzinfarkten zielen. Dagegen bevorzugt man in den Phasen III und IV klinische Endpunkte: Das sind Symptome oder Zustände, die ein Patient unmittelbar an sich wahrnehmen kann (etwa das Auftreten eines Symptoms oder die Linderung von Schmerzen) oder die Zeitdauer, die vergeht, bis ein bestimmtes Ereignis eintritt. Man unter-

scheidet zwischen „harten“ Endpunkten, die auf objektiven Fakten basieren und präzise gemessen oder erfasst werden können (etwa Laborwerte, das Auftreten einer Nebenwirkung oder eines Rezidivs), und „weichen“ Endpunkten, die weniger gut objektivierbar sind (dazu zählen beispielsweise die Merkmale „Schmerzintensivität“ oder „Lebensqualität“) [3].

In aller Regel kommen bei einer Studie mehrere Endpunkte in Betracht. So wird bei Krebserkrankungen bevorzugt die progressionsfreie Zeit oder die Zeit des Überlebens als „harter“ Endpunkt gewählt, wenngleich auch der „weiche“ Endpunkt „Lebensqualität“ relevant erscheint – insbesondere dann, wenn die Lebenserwartung der Patienten nur noch wenige Monate beträgt. In jedem Fall sollte der Studienleiter einen primären Endpunkt festlegen – das ist der Parameter, der das wichtigste Ziel der Studie am besten beschreibt. Gleichwohl verhindert diese Festlegung nicht, dass weitere, sogenannte sekundäre Endpunkte in Augenschein genommen werden.

- **Einflussgrößen:** Die wichtigste Einflussgröße bei Therapiestudien ist die Art der Therapie. Darüber hinaus bietet es sich an, weitere potenzielle Einflussparameter zu erfassen, die eventuell dazu beitragen, die primäre Zielgröße zu erklären und Unterschiede zwischen Subgruppen zu erkennen (z. B. das Alter der Patienten, Risikofaktoren oder Komorbiditäten). Insbesondere bei nicht randomisierten Studien sollte man darauf achten, Störgrößen zu erheben (sogenannte Confounder), die das Ergebnis einer Studie verzerren und zu unzulässigen Schlussfolgerungen verleiten könnten, um deren Effekte später herausrechnen zu können.
- **Studiendesign:** Studien der präklinischen Phase, der Phasen I und II werden meist als prospektive Studien ohne Vergleichsgruppe durchgeführt. Dagegen werden in Studien der Phase III (zuweilen schon in der Phase II) Patienten, die die neue Therapie erhalten, mit einer Gruppe von Patienten verglichen, die anders therapiert werden (z. B. mit der bisher verwendeten Standardtherapie, „treatment as usual“, TAU). Ein solcher Vergleich ist nicht nur bei Arzneimittelstudien vorgeschrieben, sondern erscheint auch bei anderen Behandlungsoptionen wie bei Medizinprodukten geboten. Um faire Vergleiche zu ermöglichen, sollten die Gruppen strukturgleich sein. Dies wird am ehesten durch Randomisierung, also durch die zufällige Zuteilung der Studienteilnehmer zu einer Therapiegruppe, erreicht. Deshalb werden

Phase-III-Studien im Sinne der Evidenzbasierten Medizin vorzugsweise als „Randomisierte Studie“ (RCT = randomized clinical trial) durchgeführt. Weitere Kriterien stellen die Behandlungs- und die Beobachtungsgleichheit dar: Die Patienten beider Gruppen sollten in gleicher Weise behandelt (mit Ausnahme der Therapien) und beobachtet werden. Um dies zu gewährleisten, eignen sich doppelblinde Studien, bei denen weder der Patient noch der behandelnde Arzt im Einzelfall die spezielle Therapie kennt. Allerdings sind diese Qualitätskriterien nicht in jedem Fall praktisch umsetzbar (sei es aus ethischen oder aus organisatorischen Gründen). Als Alternative zu RCTs bieten sich prospektive Beobachtungsstudien an, bei denen die Zuteilung der Patienten zu einer Therapie nach anderen Kriterien erfolgt (beispielsweise durch den behandelnden Arzt oder aufgrund der Wünsche des Patienten). Studien in der Phase IV lassen sich als randomisierte Studien oder als Anwendungsbeobachtungen durchführen. Letztere sind nicht-interventionelle Studien, die dazu dienen, Informationen über bereits zugelassene Therapien zu erheben. Auch retrospektive Studien sind denkbar, bei denen die Daten beispielsweise aus Patientenakten oder einem Register rekrutiert werden.

- **Studienpopulation:** Bei Studien der Phasen II und III werden die teilnehmenden Patienten im Allgemeinen durch vorab festgelegte Ein- und Ausschlusskriterien definiert. Während die Einschlusskriterien definieren, welche Patienten von der zu testenden Therapie profitieren könnten, beschreiben die Ausschlusskriterien, welche Patienten von der Studienteilnahme auszuschließen sind (etwa weil sie Risikofaktoren aufweisen). Je strenger die Auswahlkriterien, desto homogener ist die Studienpopulation und desto besser lässt sich der therapeutische Effekt nachweisen. In jedem Fall sollte der Studienleiter darüber nachdenken, wie viele Probanden oder Patienten in seiner Klinik oder seinem Institut in absehbarer Zeit rekrutierbar sind, oder ob eventuell eine multizentrische Studie in Erwägung gezogen werden muss. Bei Studien der Phase IV gelten keine einschränkenden Auswahlkriterien. Bei retrospektiven Studien ist die Studienpopulation durch die gegebenen Daten bereits vorgegeben. In diesen Fällen muss der Forscher evaluieren, ob diese Population für seine Fragestellung geeignet ist.

Diese Darlegungen zeigen: Bei der Planung einer Studie sind mannigfaltige Überlegungen anzustellen und weitreichende Entscheidungen zu treffen, die auch

nachfolgende Entwicklungsphasen betreffen. Eine strukturierte Vorgehensweise zahlt sich hier aus. Sie hilft, die Gedanken zu ordnen und Klarheit über das eigene Forschungsvorhaben zu gewinnen. Das Forschungsvorhaben beginnt, Struktur anzunehmen.

2.2 Chaosvorbeugung durch Kompromisse

Zur konkreten Studienplanung sollte ein Statistiker herangezogen werden. Des- sen Know-How kann sehr viel dazu beitragen, das Datenchaos in Grenzen zu halten (noch ehe die Daten erhoben werden). Dies betrifft vor allem das Studien- design und den Stichprobenumfang. Bei einer prospektiven Studie sind folgende Überlegungen anzustellen:

- **Subgruppen:** Wird nur eine Gruppe untersucht (etwa bei Phase-I-Studien) oder werden mehrere Subgruppen verglichen (wie bei Phase-III-Studien)? Weisen diese Subgruppen etwa die gleiche Gruppengröße auf? Falls nein: In welchem Verhältnis stehen die Gruppengrößen?
- **Primärer Endpunkt:** Welche Eigenschaften hat dieser Endpunkt: stetig, binär, ordinal skaliert oder quantitativ abgestuft? Wie sind die Daten verteilt? Steht eine zuverlässige und validierte Messmethodik zur Verfügung?
- **Fragestellung:** Welche Vergleiche bezüglich des primären Endpunkts sollen durchgeführt werden (etwa zwischen zwei Gruppen oder Vorher-Nachher-Vergleiche)? Soll ein Überlegenheitsnachweis erbracht werden (zum Bei- spiel, dass eine neue Therapie bezüglich ihrer Wirkung einem Placebo oder der bisher verwendeten Standardtherapie überlegen ist), oder soll ein Un- terschied nachgewiesen werden? In manchen Situationen wird nicht die Überlegenheit angestrebt: Es genügt zu zeigen, dass die neue Therapie der Vergleichstherapie nicht unterlegen ist (wenn sie dafür in anderen Bereichen Vorteile aufweist, etwa weniger Nebenwirkungen oder geringere Kosten). Bei der Untersuchung von Generika ist der Nachweis der Äquivalenz zweier Therapien gefragt. Daran entscheidet sich, welcher statistische Test geeignet und welche Fallzahl erstrebenswert ist.
- **Fallzahlplanung:** Dies ist ein bedeutender Aspekt der Planungsphase, da die erforderliche Fallzahl darüber entscheidet, wie lange die Rekrutierungs- phase dauern wird und wann mit ersten Ergebnissen zu rechnen ist. Dieser

Schritt stellt eine große Herausforderung an den Statistiker dar: Die Fallzahl sollte einerseits ausreichend hoch sein, um einen relevanten Unterschied nachweisen zu können. Andererseits sollte sie nicht zu hoch sein, da dadurch Patienten unnötigerweise belastet würden und sich die Studiendauer in die Länge ziehen würde. Beide Szenarien wären aus ethischer Sicht problematisch. Um eine adäquate Fallzahl zu ermitteln, muss der Statistiker wissen, welche Ergebnisse erwartet werden und ab welcher Größenordnung ein Effekt klinisch relevant ist.

- **Poweranalyse:** Die Situation stellt sich anders dar, wenn im Rahmen von retrospektiven Studien auf bereits vorhandenes Datenmaterial zugegriffen wird. Anstelle einer Fallzahlplanung ist eine Poweranalyse erforderlich, um herauszufinden, ob und mit welchen Einschränkungen die der Studie zugrunde liegende Fragestellung beantwortbar ist. Außerdem sollte der Statistiker die Qualität des Datenmaterials dahingehend sichten, ob die Daten korrekt und vollständig erfasst sind.
- **Bias und Confounder:** Bei einem Bias handelt es sich um einen systematischen Fehler, der das Ergebnis verzerren und zu unzulässigen Schlussfolgerungen führen kann [4]. Ein Selektionsbias entsteht, wenn die Gruppe der Studienteilnehmer nicht repräsentativ für die Zielpopulation ist. Dieser Bias ist zu befürchten, wenn nur spezielle Patienten an einer Studie teilnehmen (beispielsweise aufgrund strenger Ausschlusskriterien) oder wenn viele Patienten während der laufenden Studie ausscheiden und nicht mehr für die Endanalyse zur Verfügung stehen. In diesen Fällen wären die Ergebnisse der Studie nicht verallgemeinerbar. Ein Selektionsbias liegt auch vor, wenn zwei zu vergleichende Gruppen a priori nicht strukturgleich sind. Ein Informationsbias entsteht durch nicht korrekte, lückenhafte oder ungenaue Datenerhebung oder -erfassung. Dieser Bias kann durch Mess- oder Dokumentationsfehler bedingt sein; er kann auch durch fehlerhaft ausgefüllte Fragebögen, durch falsche Angaben bei Interviews oder durch mangelhaftes Erinnerungsvermögen von Probanden entstehen. Insbesondere retrospektive Studien, bei denen die interessierenden Ergebnisse bereits stattgefunden haben, sind anfällig für Informationsbias. – Generell ist ein Bias bei keinem Studiendesign auszuschließen. Deshalb sollte man Vorsorgemaßnahmen treffen, um dessen Einfluss zu minimieren. Bei nichtrandomisierten Studien

ist darauf zu achten, dass mögliche Confounder erfasst werden, um faire Gruppenvergleiche zu ermöglichen.

Es liegt also nicht zuletzt im Verantwortungsbereich des Statistikers, dafür Sorge zu tragen, dass Chaos bei der Datenerfassung, der Analyse und der Interpretation der Ergebnisse vermieden wird.

Bei der Studienplanung sind zuweilen Kompromisse einzugehen. Die Randomisierung gilt zwar als Goldstandard bei Therapiestudien, weil damit ein Selektionsbias vermieden wird. Dieses Design lässt sich jedoch nicht erzwingen (wenn etwa Patienten Vorbehalte äußern, an einer solchen Studie teilzunehmen). Doppelverblindung wird wegen der Beobachtungsgleichheit empfohlen; dieses Design ist jedoch nicht in jedem Fall praktikabel. Eine hohe Fallzahl ermöglicht präzise Schätzungen; andererseits bereitet diese Forderung in der klinischen Forschung mitunter erhebliche Probleme (insbesondere bei selten auftretenden Krankheiten). Als primärer Endpunkt eignet sich unter statistischen Aspekten eine stetige, normalverteilte Größe; diese ist jedoch nicht unbedingt klinisch relevant. Harte Endpunkte lassen sich präzise messen oder objektiv erfassen; andererseits sind weiche Endpunkte wie beispielsweise die Lebensqualität oft aussagekräftiger. Biomarker und andere Surrogatparameter lassen sich bereits kurz nach Therapiebeginn messen; sie stellen jedoch nur Ersatzgrößen für die eigentlich interessierenden Studienziele dar (z. B. Überlebenszeit) [2]. Homogene Patientengruppen lassen sich zwar effizienter analysieren als heterogene Gruppen. Allerdings gelten die so erhaltenen Ergebnisse nur eingeschränkt; sie sind nicht ohne Weiteres auf die Zielpopulation übertragbar. Die Studienpopulation sollte repräsentativ für die Zielpopulation sein; jedoch unterliegt die Studienteilnahme einschränkenden Bedingungen. Aus Gründen der Sicherheit mag es sinnvoll sein, Risikopatienten (z. B. ältere Menschen mit Komorbiditäten) von der Studienteilnahme auszuschließen; jedoch sind es häufig genau diese Patienten, die von der Therapie profitieren würden. Eine lang andauernde Studie hat den Vorteil, dass sich Langzeitverläufe beobachten lassen. Andererseits stellt dies hohe Anforderungen an die Patienten (z. B. häufige Wahrnehmung von Untersuchungsterminen), und es vergeht entsprechend viel Zeit, ehe valide Ergebnisse vorliegen. Die zu vergleichenden Subgruppen mögen hinreichend groß und strukturgleich sein. Es ist jedoch keineswegs gewährleistet, dass dieser Zustand so lange aufrechterhalten werden kann, bis die relevanten Endpunkte erhoben werden können.

Diese Widersprüche werden sich in den seltensten Fällen zu jedermanns Zufriedenheit auflösen lassen. Es hängt vor allem von der Phase des Entwicklungsprozesses und vom konkreten Studienziel ab, welche Einschränkungen akzeptabel erscheinen. Die jeweiligen Entscheidungen sollten vom Fachvertreter und vom Statistiker gemeinsam im Hinblick auf das eigentliche Studienziel und das Wohl der Patienten getroffen werden. Jeder an einer Studie beteiligte Wissenschaftler sollte das Seine dazu beitragen, dass die Studie gut geplant wird, und dass aus Chaos Ordnung entsteht.

2.3 Vom Datenchaos zum statistischen Modell

Der Statistiker sieht sich häufig mit einer Excel-Tabelle konfrontiert, die einen schier unübersehbaren Datenwust ohne erkennbare Zusammenhänge enthält. Diese liefert (zumindest auf den ersten Blick) keinen hohen Erkenntnisgewinn. Eine strukturierte Vorgehensweise erweist sich als nützlich, um schrittweise das Dunkel zu lichten.

- **Tabelleninhalt:** Eine Beschreibung des Tabelleninhalts trägt bereits zur Veranschaulichung bei: Die Zeilen und Spalten einer Tabelle repräsentieren die Studienteilnehmer bzw. die erfassten Merkmale; einzelne Tabellenblätter könnten sich auf aufeinanderfolgende Untersuchungszeitpunkte beziehen. Ein kritischer Blick auf die Excel-Tabelle lässt erahnen, ob die Daten möglichst vollständig oder eher lückenhaft erfasst sind, ob gravierende Fehler erkennbar sind und ob unnütze Daten, die für die Analyse nicht relevant sind, enthalten sind. Außerdem sollte geprüft werden, ob alle relevanten Ziel- und Einflussgrößen erfasst sind.
- **Plausibilitätsprüfung:** Zur statistischen Analyse empfiehlt sich die Verwendung einer validierten Statistiksoftware (z. B. R, SAS oder SPSS). Um eine hohe Datenqualität zu gewährleisten, sollten nach dem Datenimport Plausibilitätsprüfungen durchgeführt werden. Ein Blick auf die Häufigkeitsverteilung eines jeden Merkmals zeigt, ob die aufgelisteten Werte oder Klassifikationen realistisch sind. Grobe Fehler bei der Dateneingabe werden dabei offensichtlich.
- **Ausreißer:** Dies sind extrem hohe oder extrem niedrige Werte eines quantitativen Merkmals, bei denen fraglich ist, ob sie unter denselben Bedingun-

gen wie die anderen Werte der Datenreihe erhoben wurden. Durch einen Blick auf das Minimum oder Maximum sind Ausreißer leicht auffindbar. Die Einstufung eines Wertes als Ausreißer ist in erster Linie inhaltlich motiviert. Zunächst ist – am besten mit einem klinischen Fachvertreter – zu klären, ob es sich um einen falschen Eintrag (aufgrund eines Mess-, Dokumentations- oder Rechenfehlers), um eine pathologische Besonderheit oder um eine physiologische Rarität handelt. Fehlerhafte Einträge, die sich nicht korrigieren lassen, sind freilich von der Analyse auszuschließen. Wenn sachlogische Gründe dies erfordern (etwa bei Non-Konformität mit den Ein- und Ausschlusskriterien), müssen gar die dazugehörigen Datensätze komplett eliminiert werden. Ansonsten sind Ausreißer bei der statistischen Analyse durch die Wahl geeigneter Methoden und bei der Interpretation der Ergebnisse in angemessener Weise zu berücksichtigen.

- **Deskriptive Statistik:** Mit diesen Methoden werden charakteristische Eigenschaften eines einzelnen Merkmals beschrieben. Wenn zwei Gruppen verglichen werden, sollten beide Subgruppen separat beschrieben werden. Bei qualitativen Merkmalen werden absolute und relative Häufigkeiten ermittelt; für quantitative Variablen lassen sich Lagemaße, Streuungsmaße und Formmaße berechnen. Mit einem Assoziationsmaß (z. B. Korrelationskoeffizient oder *Odds Ratio*) lässt sich die Stärke des Zusammenhangs zwischen zwei Merkmalen quantifizieren. Auch graphische Darstellungen (Balkendiagramme, Histogramme oder *Box-and-Whisker-Plots*) eignen sich, um Besonderheiten einer Datenreihe aufzuspüren. Das geübte Auge eines Statistikers wird wertvolle Informationen aus diesen Darstellungen extrahieren und erkennen, inwieweit diese Ergebnisse vielversprechend in Hinblick auf das Studienziel sind und worauf er in den folgenden Analyseschritten zu achten hat. Das anfängliche Chaos hat sich zu diesem Zeitpunkt bereits erkennbar reduziert.
- **Gruppenvergleiche:** Der Vergleich ist ein wichtiges Element in der klinischen und der epidemiologischen Forschung: Bei Therapiestudien werden meist zwei Therapieformen, bei Fall-Kontroll-Studien erkrankte Patienten mit gesunden Kontrollen verglichen. Dichotomisierungen sind allseits beliebt: Sie machen das Leben einfach und bringen Struktur ins Chaos. Der Statistiker wird sich bemühen, vorhandene (oder nicht vorhandene) Unter-

schiede zwischen den Gruppen mittels statistischer Tests zu belegen. So kann er beurteilen, ob die beiden Gruppen zu Studienbeginn strukturgleich sind und ob sich die Gruppen bezüglich des Endpunkts unterscheiden.

- **Interpretation der Ergebnisse:** Der Statistiker ist vor die Aufgabe gestellt, aus der Vielzahl von statistischen Testverfahren jeweils einen adäquaten Test auszuwählen und diese Wahl zu begründen. Das Ergebnis eines jeden Tests ist der sogenannte p-Wert. Dieser Wert quantifiziert den sogenannten α -Fehler (auch Fehler 1. Art genannt). Dies ist Wahrscheinlichkeit dafür, dass das erhaltene Testergebnis allein aufgrund des Zufalls zustande gekommen ist. Gemeinhin bezeichnet man ein Testergebnis mit einem p-Wert, der geringer ist als 0,05, als „statistisch signifikant“. Zusätzlich zum p-Wert sollte eine Effektgröße angegeben werden (etwa die Differenz zwischen zwei Mittelwerten oder eine *Odds Ratio*), am besten zusammen mit einem Konfidenzintervall. Nur mit diesen Angaben lässt sich beurteilen, ob ein statistisch signifikantes Ergebnis klinisch relevant ist und wie verlässlich die Schätzung ist. Ein nicht-signifikantes Ergebnis bedeutet keineswegs, dass definitiv kein Unterschied zwischen zwei Gruppen vorhanden ist. Es könnte auch aufgrund eines zu geringen Stichprobenumfangs zustande kommen (dies wäre ein β -Fehler oder Fehler 2. Art). Auch in diesem Fall sind der p-Wert und ein Konfidenzintervall nützlich, um das Testergebnis angemessen zu interpretieren.
- **Analyse des Endpunkts:** Meist werden bei einer Therapiestudie neben der Therapieform weitere potenzielle Einflussgrößen erhoben. Es ist nützlich zu evaluieren, ob und in welchem Maße jede dieser Einflussgrößen mit dem Endpunkt assoziiert ist. So lässt sich einschätzen, welche Einflussgrößen wichtig oder weniger wichtig sind, und bei welchen am ehesten Kausalitäten zu vermuten sind. Häufig wird sich zeigen, dass eine Zielgröße nicht nur mit einer, sondern mit mehreren Einflussgrößen statistisch assoziiert ist. Dann lohnt es sich, eine multiple Analyse durchzuführen, bei der simultan mehrere Einflussgrößen evaluiert werden [11]. Ein multiples Verfahren weist gegenüber einfachen statistischen Tests mehrere Vorteile auf: Individuelle Charakteristika der Studienteilnehmer können im Rahmen eines multiplen Modells angemessen berücksichtigt werden. Bei nicht randomi-

sierten Studien helfen multiple Methoden, nach Confoundern zu adjustieren und Verzerrungen zu vermeiden [4,11].

Die Konstruktion eines adäquaten statistischen Modells verlangt sehr viel Know-How und detailliertes Fachwissen. Anders als man es bei Lösungen von mathematischen Problemstellungen gemeinhin erwartet, ist das finale Modell keineswegs eindeutig bestimmt. Auch bei einer gut geplanten Studie ergeben sich an vielen Stellen der statistischen Analyse Fragen, die zu Kompromissen zwingen. Das beginnt bei der Wahl der Testverfahren: Einerseits werden Tests mit einer hohen Power bevorzugt; diese stellen jedoch Anforderungen an die Daten, die mitunter nicht gegeben oder nicht überprüfbar sind. Andere Fragen betreffen den Umgang mit fehlenden Werten: Sollte man diese durch Imputationsverfahren schätzen (und damit Ungenauigkeiten in Kauf nehmen), oder sollte man die jeweiligen Datensätze bei der Analyse nicht berücksichtigen (was einen Verlust an Power mit sich bringt)? Wie handhabt man Ausreißer? Welche Variablen sollten in ein multiples Modell aufgenommen werden [11]? Was besagen die Ergebnisse der statistischen Analyse bezüglich Kausalitäten? Besonders bei dieser Frage sollte man sehr vorsichtig argumentieren, um nicht zu irreführenden Schlussfolgerungen zu gelangen.

Nichtsdestotrotz bleibt festzuhalten: Leistungsstarke Rechner und effiziente Analysemethoden ermöglichen es, basierend auf den Rohdaten eine Regressionsgleichung zu erstellen, die die interessierende Zielgröße (z. B. die Wirkung einer Therapie oder die Wahrscheinlichkeit für das Eintreten eines Ereignisses) in Abhängigkeit von mehreren relevanten Einflussfaktoren in präziser Formelsprache beschreibt. Damit lassen sich Phänomene erklären, Zusammenhänge erkennen, Kausalitäten vermuten und Prognosen wagen. Eine solche Gleichung ist faszinierend! Sie versinnbildlicht: Aus dem Chaos ist ein strukturiertes Modell entstanden, das die Wirklichkeit zwar nicht vollständig, aber zumindest teilweise abbildet. Mehr noch: Das unstrukturierte Gedankenchaos im Hirn eines Forschers und das Chaos der Rohdaten haben sich aufgelöst und Geheimnisse preisgegeben. Daraus sind neue Erkenntnisse hervorgegangen.

3 Anwendungsbeispiele

3.1 Historische Beispiele

Die in Abschnitt 2 dargelegten Empfehlungen, die dazu dienen, ein Gedankenchaos zu ordnen und ein Datenchaos zu strukturieren, sind in der klinischen und epidemiologischen Forschung noch nicht allzu lange etabliert. Heute herrscht zwar allgemeiner Konsens darüber, dass statistische Analysen für die medizinische Forschung unverzichtbar sind. Jedoch hatten Ärzte und Patienten bis weit ins 20. Jahrhundert hinein massive Vorbehalte bezüglich Studien in der klinischen Forschung. Erst im Jahre 1964 wurden auf der 18. Generalversammlung des Weltärztebundes in Helsinki Richtlinien für die Forschung am Menschen als „Deklaration von Helsinki“ erarbeitet. Diese Deklaration wurde seither mehrfach revidiert und neuen Erkenntnissen angepasst; die letzte Revision wurde im Jahre 2013 in Fortaleza (Brasilien) verabschiedet. Die technischen Voraussetzungen, die mathematischen Verfahren und adäquate Softwareprodukte, die erforderlich sind, um große Datenmengen effizient zu bearbeiten und aus einem Datenchaos ein statistisches Modell zu generieren, entstanden erst mit dem Aufkommen leistungsfähiger Computer und benutzerfreundlicher Software in den 1980er Jahren.

Dennoch haben auch in früheren Zeiten, lange vor der Deklaration von Helsinki und dem EDV-Zeitalter, Wissenschaftler versucht, Daten zu generieren und auszuwerten, um Antworten auf Fragen zu finden, die sie allein mit ihrem Fachwissen nicht beantworten konnten. Als Beispiel seien die Forschungsarbeiten des Gynäkologen *Ignaz Philipp Semmelweis* (1818–1865) erwähnt. Semmelweis arbeitete um die Mitte des 19. Jahrhunderts in einer geburtshilflichen Abteilung des Wiener Allgemeinen Krankenhauses. In dieser Abteilung betreuten Ärzte und Medizinstudenten die Frauen bei der Entbindung, während in einer zweiten Abteilung ausschließlich Hebammen als Geburtshelferinnen fungierten. In Semmelweis' Abteilung trat ein gravierendes Problem zutage: Dort starben sehr viele Mütter an Kindbettfieber, wesentlich mehr als in der Abteilung, die von Hebammen geführt wurde. Die Fachwelt stand vor einem Rätsel. Es wurden zahlreiche, meist abstruse Theorien aufgestellt, als mögliche Gründe für die hohe Mortalität wurden diskutiert: Stau des Wochenflusses; Blutunreinheiten, die während der Schwangerschaft entstehen; Wettereinflüsse; Angst und Scham der Frauen; Verunreinigung der Muttermilch oder verunreinigte Luft [8]. Freilich waren all diese Theorien rein spekulativ und für einen kritischen Forscher nicht überzeugend.

Semmelweis war sehr daran gelegen, die Krankheitsursachen zu ergründen. So seziierte er, seine Kollegen und Studenten Leichen von Frauen, die der Krankheit zum Opfer gefallen waren. Dabei ging Semmelweis strukturiert vor: Er notierte seine Befunde und versuchte, Muster zu erkennen. Die Mortalitätsrate sank jedoch nicht. Das Phänomen blieb mysteriös, auch nachdem Tausende von Leichen untersucht worden waren.

Semmelweis entschloss sich, weitere Informationen einzuholen. Er setzte er sich mit Kollegen in Verbindung, um deren Erfahrungen einzuholen. Er las sämtliche Bücher und Zeitschriften, die sich dieser Thematik widmeten und setzte sich mit den Ansichten seiner Kollegen kritisch auseinander. Mit diesem Hintergrundwissen stellte er alle bis dahin bekannten Fakten zusammen. Dabei konnte er sich auch auf seine eigenen, mannigfaltigen Beobachtungen am Krankenbett und am Seziertisch und auf seine sorgfältigen Dokumentationen stützen. Über einen Zeitraum von mehreren Jahren hatte er festgehalten, wie viele Frauen an welchen Tagen in seiner Abteilung entbunden wurden, wie viele erkrankten und wie viele verstarben. Er stellte fest [8]:

- In beiden Abteilungen des Wiener Krankenhauses wurden pro Jahr etwa gleich viele Entbindungen verzeichnet (jeweils 3000–3500). Die Anzahl der Todesfälle durch Kindbettfieber belief sich auf 600–680 in der Abteilung, in der Ärzte arbeiteten. Dagegen verstarben in der Hebammenabteilung nur etwa 60 Frauen pro Jahr.
- Die Sterblichkeitsraten bei Hausgeburten und bei privat praktizierenden Gynäkologen waren wesentlich geringer als die Raten in Krankenhäusern, in denen Ärzte bei der Entbindung zugegen waren. Dies galt nicht nur in Wien, sondern auch in anderen Städten, die geburtshilfliche Abteilungen in Krankenhäusern unterhielten (u. a. in Paris, London oder Dresden).
- Das Phänomen war keineswegs neu. Bereits im Jahre 1746 (also mehr als 100 Jahre vor Semmelweis' Zeit) wurde über eine Kindbettfieber-Epidemie am Pariser Hospital Hôtel-Dieu berichtet.
- Unter dem ehemaligen Leiter der Wiener Geburtsklinik Dr. Johann Boër (1751–1835), der sanfte Methoden bei der Geburtshilfe empfahl und nur im Notfall Instrumente einsetzte, lag die durch Kindbettfieber bedingte Mortalität über 30 Jahre lang bei etwa einem Prozent. Erst unter Boërs

Nachfolger Johann Klein (1788–1856), der auch Semmelweis' Chef war, stieg die Mortalität sprunghaft an.

- Je traumatischer die Geburt war, desto höher war die Wahrscheinlichkeit zu erkranken.
- Die Erkrankungsrate (Inzidenz) war unabhängig vom Wetter.
- Mit einer Frau erkrankte meist auch deren Kind.

Damit konnte Semmelweis das Phänomen der hohen Mortalität zwar nicht erklären. Er konnte aber einige der gängigen Theorien widerlegen und kam seinem Ziel, die Ursache des Kindbettfiebers zu ergründen, ein wenig näher. Letztlich kam ihm der Zufall zu Hilfe: Nachdem sich sein Kollege Jakob Kolletschka (1803–1847) beim Sezieren einer Leiche mit dem Messer verletzt hatte und an Wundfieber erkrankte, erkannte Semmelweis die Parallelität der beiden Krankheitsbilder. Daraus schloss er messerscharf, dass den Frauen durch die Ärzte und Studenten, die in seiner Abteilung arbeiteten, „Kadaverteilchen“ übertragen wurden, die mit bloßem Auge nicht sichtbar waren. Als Konsequenz ordnete an, dass sich fortan alle Ärzte, Studenten und Hebammen mit Chlorkalk die Hände desinfizieren sollten, ehe sie den Frauen bei der Entbindung beistanden. Die Mortalität sank daraufhin dramatisch auf etwa 1,3 % im Jahre 1848. Semmelweis war es gelungen, durch strukturiertes Vorgehen und durch statistische Analysen die Krankheitsursache ausfindig zu machen und erfolgreich zu bekämpfen. Er gilt heute als einer der Väter der Evidenzbasierten Medizin [1].

Es lassen sich weitere Beispiele für medizinische Studien aus der Zeit des 19. Jahrhunderts aufführen [10]. Als im Jahre 1854 im Londoner Stadtteil Soho eine Cholera-Epidemie wütete, führte der Chirurg *John Snow* (1813–1858) eine systematische Untersuchung durch, indem er zahlreiche Erkrankte und Nicht-Erkrankte befragte. Mittels diesen Zwei-Gruppen-Vergleichs konnte er nachweisen, dass die Krankheit auf verunreinigtes Trinkwasser, das einem bestimmten Brunnen entnommen wurde, zurückzuführen war. Nachdem er den Brunnen außer Betrieb gesetzt hatte, nahm die Zahl der Neuerkrankungen drastisch ab. Damit konnte Snow die damals gängige Miasmen-Theorie widerlegen, adäquate Maßnahmen treffen und auf diese Weise viele Menschenleben retten. – Der französische Arzt *Pierre Alexandre Louis* (1787–1872) führte am Pariser Krankenhaus „Hôtel-Dieu“ eine retrospektive Studie durch, indem er zwei Gruppen von lungenkranken

Patienten verglich: Die Patienten der einen Gruppe waren zur Ader gelassen worden, während die anderen mit Bettruhe therapiert worden waren. Louis konnte nachweisen, dass die Behandlung mit Aderlass nutzlos oder gar schädlich war. Heute gilt er als der Gründer der „numerischen Methode“. – Zuletzt sei der Augustinermönch *Johann Gregor Mendel* (1822–1884) genannt, der seine Vererbungsgesetze mit statistischen Methoden verifizierte.

Leider war es diesen Forschern nicht vergönnt, den Ruhm für ihre Entdeckungen zu Lebzeiten zu ernten. Von ihren Kollegen, die sich in ihrer Ehre verletzt fühlten, ernteten sie vielfach Hohn und Spott. Die Methodik, Daten zu analysieren, um zu neuem Wissen zu gelangen, stieß überwiegend auf Ablehnung. Dennoch haben sich ihre Ideen langfristig durchgesetzt und die klinische Praxis maßgeblich beeinflusst. Auch wenn damals systematische Studien unbekannt waren und die angewandten statistischen Methoden trivial erscheinen, sollten wir anerkennend festhalten: Sie waren Pioniere auf dem Gebiet der klinischen Forschung. Es gelang ihnen durch eine strukturierte Vorgehensweise, ihr Gedankenchaos zu sortieren, unsinnige Theorien zu widerlegen, Auffälligkeiten in ihrem Zahlenmaterial zu entdecken und auf diese Weise, teilweise gegen heftige Widerstände, der Wissenschaft und zahllosen Patienten einen wertvollen Dienst zu erweisen.

3.2 Aktuelle Studien

Bleiben wir bei Entbindungen: Im Jahre 2016 wurden die Ergebnisse einer randomisierten Studie von Kirscht et al. publiziert, in der die Auswirkungen von mechanischer Dilatation der Zervix bei Frauen, die durch einen Kaiserschnitt entbunden wurden, untersucht wurde [5]. Insgesamt wurden 447 Frauen zufällig einer Gruppe (mit oder ohne Dilatation) zugewiesen. Die primäre Zielgröße war das Auftreten einer postpartalen Hämorrhagie (Blutung) innerhalb von sechs Wochen nach der Geburt. Dieser Endpunkt wurde gewählt, weil die postpartale Hämorrhagie eine lebensbedrohliche Komplikation darstellt, und weil noch keine randomisierte Studie zu dieser Thematik vorlag. Vergleiche der beiden Gruppen bezüglich demographischer Merkmale und Baseline-Charakteristika (Alter, BMI, Schwangerschaftsdauer, Gravidität, Parität, Mehrlingsschwangerschaft, Geburtsgewicht des Kindes u. a.) zeigten, dass die zu vergleichenden Gruppen weitgehend strukturgleich waren – wie man das bei einer randomisierten Studie dieses Umfangs erwartet –, und dass das Endergebnis nicht durch Confounder verzerrt sein würde.

Mit einem einfachen Gruppenvergleich ergab sich kein signifikanter Unterschied bezüglich des primären Endpunkts zwischen den Gruppen mit und ohne Dilatation: Die Inzidenzen betragen 2,4 % bzw. 1,2 % ($p = 0,479$). Allerdings gab es einen hoch signifikanten Unterschied bezüglich des Anteils der Frauen, bei denen Teile der Schwangerschaftsanlage im Uterus verblieben (0 % versus 6,2 %; $p < 0,0001$). Bezüglich weiterer sekundärer Endpunkte (Operationsdauer, Notwendigkeit einer Bluttransfusion, Auftreten von Kindbettfieber u. a.) waren keine signifikanten Unterschiede nachweisbar. Bei der Analyse des primären Endpunkts ergab sich, dass ein erhöhter BMI und eine Mehrlingsschwangerschaft mit einem höheren Risiko für das Auftreten einer Hämorrhagie assoziiert waren. Schließlich wurde ein multiples Modell erstellt, das es ermöglicht, die Wahrscheinlichkeit für das Auftreten einer postpartalen Hämorrhagie in Abhängigkeit der Einflussfaktoren Dilatation (ja/nein), BMI und Mehrlingsschwangerschaft (ja/nein) zu schätzen.

Als zweites Beispiel betrachten wir eine Studie der Adipositaschirurgie [7], bei der zwei Operationstechniken verglichen wurden, die eine Gewichtsreduktion bewirken sollten: ein Magenbypass, bei dem der Magen unterhalb des Mageneingangs abgetrennt wird, und eine Schlauchmagenbildung, bei der das Organ irreversibel verkleinert wird. Lange Zeit wurde der Nutzen dieser Techniken kontrovers diskutiert. Es herrschte Unklarheit, ob eine und gegebenenfalls welche Operationstechnik der anderen überlegen war. In diesem Fall schien eine randomisierte Studie aus ethischen Gründen problematisch zu sein, da die Entscheidung für eine bestimmte OP-Technik der behandelnde Chirurg üblicherweise zusammen mit dem Patienten traf. Also entschloss man sich, eine prospektive Beobachtungsstudie durchzuführen. 127 partizipierende Patienten erhielten einen Bypass, nur 46 Patienten einen Schlauchmagen. Die zu vergleichenden Gruppen unterschieden sich a priori bezüglich ihres BMIs: Bei den Bypass-Patienten betrug er im Mittel ($45,6 \pm 5,7$) kg/m^2 , während er bei den Patienten, die einen Schlauchmagen erhielten, mit durchschnittlich ($55,9 \pm 7,8$) kg/m^2 wesentlich höher war. Als primäre Zielgröße wurde der Exzess Weight Loss (EWL) gewählt. Diese Variable quantifiziert, um wieviel Prozent das Übergewicht durch die chirurgische Maßnahme ein Jahr nach dem Eingriff reduziert war. Beim direkten Vergleich ergab sich, dass der Bypass eine höhere Wirksamkeit aufwies ($(62,9 \pm 18) \%$ versus $(52,3 \pm 15) \%$; $p = 0,0024$). Dieses Ergebnis war jedoch wegen mangelnder Strukturgleichheit mit einem Selektionsbias behaftet. Um dies zu kompensieren, wurde eine Kovarianzanalyse als multiples Verfahren durchgeführt, in der nach dem Baseline-BMI

adjustiert wurde. Demnach beeinflusste zwar der BMI das Endergebnis ($p = 0,01$); der Unterschied bezüglich des EWL war aber nicht mehr signifikant. Am Ende wurde ein statistisches Modell entwickelt, womit der zu erwartende Wert für den EWL im Einzelfall mittels einer linearen Regressionsgleichung prognostiziert werden kann. Darüber hinaus wurden noch zahlreiche sekundäre Endpunkte analysiert (z. B. die Änderung von Stoffwechselfparametern). Dabei zeigte sich, dass der BMI vor der Operation in jedem Fall die wichtigste Einflussgröße war.

Es ist nicht anzunehmen, dass eine dieser beiden Studien die klinische Praxis nachhaltig beeinflusst hat. Trotzdem führten sie zu neuen Erkenntnissen.

4 Schlussfolgerungen

Der Vergleich von Studien aus zwei Jahrhunderten macht deutlich: Die medizinische Forschung hat sich in den vergangenen 200 Jahren dramatisch entwickelt [9]. Semmelweis und Louis befassten sich mit Krankheitsbildern, die zu ihrer Zeit lebensbedrohlich waren, weil keine wirksame Therapie zur Verfügung stand. Also war man bemüht, Risikofaktoren zu evaluieren (Semmelweis) oder die Nutzlosigkeit einer Maßnahme nachzuweisen (Louis), um das Entstehen bzw. die Progression einer Krankheit zu verhindern. Die beobachteten Effekte waren so überzeugend, dass es keiner aufwändigen statistischen Analysen bedurfte, um diese abzusichern. Die kausalen Zusammenhänge waren offensichtlich, auch ohne ein ausgeklügeltes Studiendesign. Um ethische Belange der Patienten machte sich damals niemand Gedanken. – Anders die heutige Situation: Meist steht ein Krankheitsbild im Fokus des Interesses, das nicht lebensbedrohlich ist und für das eine effiziente Therapie bereits verfügbar ist (wenngleich auch in der heutigen Zeit Krankheiten erforscht werden, die das weitere Leben der Patienten massiv beeinträchtigen, etwa in der Onkologie oder in der Neonatologie). Die Voraussetzungen zum Durchführen einer Therapiestudie sind in mehreren Gesetzesvorlagen verankert, unter anderem im Arzneimittelgesetz (AMG), im Medizinproduktegesetz (MPG) und in den Leitlinien der GCP (Good Clinical Practice). Vor Studienbeginn ist ein Votum der zuständigen Ethikkommission einzuholen. Dadurch soll gewährleistet werden, dass dem Wohle des Patienten absolute Priorität eingeräumt wird.

Vieles ist einfacher geworden. Eine Literaturrecherche und der Zugriff auf relevante Informationen ist dank des Internets quasi zu jeder Zeit und an je-

dem beliebigen Ort möglich. Es existieren eine Fülle von Fachzeitschriften und entsprechend viele Publikationen, sodass sich jeder Forscher umfassend über den aktuellen Stand seines Fachgebiets informieren kann. Die Analyse großer Datenmengen stellt – zumindest in technischer Hinsicht – keine allzu große Herausforderung dar. Anderes ist jedoch komplizierter geworden: Heutzutage werden ungleich mehr Studien durchgeführt als in früheren Zeiten, woraus entsprechend viele Publikationen resultieren. Die Datenerfassung ist weniger aufwändig, was mitunter zu einer enormen Datenmenge führt. Da fällt es mitunter schwer, die Spreu vom Weizen zu trennen.

Jeder Fortschritt entsteht aus einem Chaos – unabhängig von Zeit und Ort. Um dieses Chaos aufzulösen, bedarf es eines kritischen Geistes und einer strukturierten Vorgehensweise. Forscher – egal in welchem Jahrhundert sie leben – sollten sich zunächst umfassend über den aktuellen Stand ihres Forschungsgebietes informieren. Dazu gehört die Fähigkeit zu beurteilen, welche Erkenntnisse relevant und nachvollziehbar sind. Semmelweis musste evaluieren, welche der damals sich im Umlauf befindlichen Theorien zur Entstehung von Kindbettfieber sinnvoll und welche zu verwerfen waren. In der heutigen Zeit steht ein Forscher vor anderen Herausforderungen: Er muss aus der Fülle von Publikationen zu seinem Fachgebiet die wirklich relevanten aussortieren, lesen und bewerten. Bereits Semmelweis erkannte die Wichtigkeit von systematischen Untersuchungen, ohne die keine aussagekräftigen Informationen zu gewinnen sind. Daran hat sich bis heute nichts geändert. Während sich in früheren Zeiten die Datenanalyse im Wesentlichen auf das zeitraubende Zählen von Ereignissen und Vergleichen von Häufigkeiten beschränkte, stehen heute eine Fülle von effizienten Analysetechniken zur Verfügung, die in Sekundenschnelle Ergebnisse liefern. Dennoch ist es keineswegs trivial, eine schier unüberschaubare Datenmenge zu strukturieren und aus der Vielzahl von statistischen Techniken geeignete Verfahren auszuwählen. Mitunter verleitet dies dazu, ohne Bedacht eine Vielzahl von Tests durchzuführen in der Hoffnung, irgendein signifikantes Ergebnis zu erhalten. Dies ist jedoch keine empfehlenswerte Strategie, um dem Chaos zu entfliehen!

Der Umgang mit Daten lässt sich zwischen zwei Polen einordnen: Auf der einen Seite das Chaotische, Unstrukturierte, Unüberschaubare, Unsichere – versinnbildlicht durch die Rohdaten. Auf der anderen Seite eine Struktur, dargestellt durch ein reduktionistisches Modell, das die Komplexität des untersuchten Sachverhalts in mathematisch präziser Formelsprache beschreibt. Ein solches Modell

zu generieren, erfordert Neugierde und Leidenschaft und den unbedingten Willen des Forschenden, den Dingen auf den Grund zu gehen.

Das Reduzieren von Chaos besteht im Erkennen von Regelmäßigkeiten. Dafür sind eine gut durchdachte Planung, sorgfältig durchgeführte Untersuchungen und eine effiziente Datenanalyse unerlässlich. In letzter Zeit spielen zunehmend häufiger Begriffe wie *Big Data*, *Data Science* oder Künstliche Intelligenz (KI), die vielfach synonym verwendet werden, eine wichtige Rolle. Mit diesen Techniken sind hohe Erwartungen verknüpft: Die immensen Datenmengen bergen ein ungeheures Potential in sich. Möglicherweise lassen sich daraus Algorithmen herleiten, die bei der Früherkennung von Krankheiten, zur Vorhersage von akuten Ereignissen und bei der Entwicklung diagnostischer Tests dereinst eingesetzt werden könnten. Andererseits besteht gerade mit Big Data die Gefahr, im Chaos zu versinken!

Allerdings kann das Verharren an neuen Erkenntnissen weitere Forschung behindern [6]. Es birgt die Gefahr in sich, dass man – ohne die Besonderheiten des Einzelfalls gebührend zu berücksichtigen – Entscheidungen trifft, die sich im Nachhinein als wenig sinnvoll erweisen, oder dass man andere Aspekte außer Acht lässt. Wenn beispielsweise in einer randomisierten Studie die Überlegenheit eines neu entwickelten Arzneimittels im Vergleich zu einer Standardtherapie nachgewiesen wurde, ist man als behandelnder Arzt mitunter versucht, jedem Patienten die vermeintlich wirkungsvollere Therapie zu verordnen ohne einen Gedanken daran zu verschwenden, ob sie im konkreten Einzelfall wirklich erfolgversprechend ist. Man konzentriert sich auf eine einzige Zielgröße – etwa die Wirksamkeit – und lässt besondere Charakteristiken eines Patienten oder andere Eigenschaften der Therapie (z. B. Nebenwirkungen) außer Acht. Damit wird der Patient nicht mehr als ein Individuum wahrgenommen, sondern als eine Beobachtungseinheit, die durch eine mathematische Gleichung berechenbar ist.

Chaos und Struktur – das ist kein Widerspruch! Chaos kann Verwirrung stiften und gleichzeitig die Fantasie anregen. Strukturen können einerseits einengen und andererseits Sicherheit bieten. Fazit: Die Mischung macht's! Jeder Mensch (und jeder Wissenschaftler) braucht ein gewisses Maß an Chaos – sonst wäre das Leben langweilig, und jegliche Kreativität wäre im Keime erstickt. Im Chaos zu verharren, ist jedoch auf Dauer unbefriedigend. Deshalb werden Maßnahmen ergriffen, um das Chaos aufzulösen und ein Modell zu generieren, das die komplexen Zusammenhänge zumindest partiell beschreibt. Aber ein solches Modell ist

niemals perfekt! Deshalb verlangt auch dieser Zustand nach einiger Zeit Anpassung an neue Gegebenheiten. Immer wieder werden sich neue Herausforderungen ergeben – und der Kreislauf beginnt von vorn. Der Philosoph Manfred Hinrich (1926–2015) hat dies mit folgendem Aphorismus treffend zum Ausdruck gebracht: „Chaosbewältigung ist das halbe Leben, Chaosschaffen die andere Hälfte“.

Literatur

1. Aigner F, Grafenholder D: Leben retten, mit Chlorkalk und Statistik. <https://www.naklar.at/content/features/semmelweis/>
2. Atkinton AJ Jr, Colburn WA, De Gruttola VG, DeMets DL, Dwoning GJ, Hoth DF, Oates JA, Peck CC, Schooley RT, Spilker BA, Woodcock J, Zeger SL: Biomarkers and Surrogate Endpoints: preferred definitions and conceptual framework. *Clin Pharmacol Ther* 69 (3): 89–95; 2001
3. Gierschik P: Die Phaseneinteilung klinischer Studien. In: Lenk C, Duttge G, Fange-rau H (Hrsg.): *Handbuch Ethik und Recht der Forschung am Menschen*. Springer-Verlag 2014
4. Hammer GP, Prel du JP, Blettner M: Vermeidung verzerrter Ergebnisse in Beobachtungsstudien. *Dtsch Arztebl Int* 1061 (41): 664–668; 2009
5. Kirscht J, Weiß C, Nickol J, Berlit S, Tuschy B, Hoch B, Trebin AV, Große-Steffen T, Sütterlin M, Kehl S: Dilatation or no dilatation of the cervix during cesarean section (Dondi Trial): a randomized clinical trial. *Arch Gynecol Obstet* 295: 39–43; 2017
6. Kriz J: Chaos, Angst und Ordnung. *Ärztliche Praxis. Neurologie Psychiatrie* 1; 24–27; 1999
7. Otto M, Elrefai M, Krammer J, Weiß C, Kienle P, Hasenberg T: Sleeve Gastrectomy and Roux-en-Y Gastric Bypass Lead to Comparable Changes in Body Composition after Adjustment for Initial Body Mass Index. *Obes Surg* 26(3): 479–485; 2016
8. Nuland SB: *Ignaz Semmelweis. Arzt und großer Entdecker*. Piper Verlag München 2008
9. Tröhler U: Die therapeutische „Erfahrung“ – Geschichte ihrer Bewertung zwischen subjektiv sicherem Wissen und objektiv wahrscheinlichen Kenntnissen. In: Köbberling J: *Die Wissenschaft in der Medizin. Selbstverständnis und Stellenwert in der Gesellschaft*. Schattauer-Verlag 1992

10. Weiß C: Entwicklung der Medizinischen Statistik in Deutschland. Der lange Weg dahin. GMSD Med Inform Biom Epidemiol; 2005
11. Weiß C: Regressionsanalysen. In: Basiswissen Medizinische Statistik, Springer-Verlag, 7. Auflage 2019

Über die Autorin

Christel Weiß ist Professorin für Biomathematik und Epidemiologie an der Medizinischen Fakultät Mannheim der Universität Heidelberg. Zu ihren Tätigkeiten zählen Lehrveranstaltungen für Studierende der Medizin und Masterkurs-Absolventen, das Durchführen von Seminaren sowie die Beratung von Ärzten, wissenschaftlichen Mitarbeitern und Doktoranden bei der Planung und Durchführung von klinischen und epidemiologischen Studien. Frau Weiß ist Autorin des Lehrbuchs „Basiswissen Medizinische Statistik“ (erschieden im Springer-Verlag, 7. Auflage), des Ratgebers „Promotion. Die medizinische Doktorarbeit – von der Themensuche bis zur Dissertation“ (zusammen mit Prof. Dr. Axel Bauer, erschienen im Thieme-Verlag, 4. Auflage) sowie Koautorin zahlreicher Papers und Buchbeiträge.

Korrespondenz:

Prof. Dr. Christel Weiß

Medizinische Fakultät Mannheim der Universität Heidelberg

Abteilung für Medizinische Statistik und Biomathematik

Theodor-Kutzer-Ufer 1

68167 Mannheim, Germany

E-Mail: Christel.Weiss@medma.uni-heidelberg.de