Gefäßchirurgie 2017 · 22:412–419
<https://doi.org/10.1007/s00772-017-0306-3>
 Online publiziert: 8. September 2017
 © Der/die Autor(en) 2017. Dieser Artikel ist eine Open-Access-Publikation.

U. Mansmann

Institut für medizinische Informationsverarbeitung, Biometrie und Epidemiologie, Ludwig Maximilians Universität München, München, Deutschland

Statistische Modelle in der Gefäßchirurgie

Die Validität statistischer Methoden, der verwendeten Modelle und der implizit damit verbundenen Voraussetzungen bestimmt die Glaubwürdigkeit und Interpretierbarkeit von Studienergebnissen aus Versorgungs- wie patientenorientierter Forschung. Ändern sich diese Ergebnisse, wenn die Daten mit anderen Verfahren ausgewertet werden? Wurde die Studie mit dem optimalen statistischen Ansatz ausgewertet? Bleiben die Ergebnisse robust, wenn die Daten von den durch die statistischen Methoden implizierten Voraussetzungen abweichen? Der Artikel soll helfen, sich solchen wichtigen Fragen zu stellen.

Einleitung

Statistische Modelle sind mathematische Konstrukte, die Annahmen über Daten einer Stichprobe formalisieren. Noch allgemeiner formalisieren sie Eigenschaften einer Population, aus der die Datenstichprobe gezogen wurde. Statistische Modelle repräsentieren in einer recht idealisierten Form datengenerierenden Prozesse und Wahrscheinlichkeitsverteilungen. Damit approximieren sie die Verteilungen, aus denen beobachtete Stichproben erzeugt wurden. Auf statistischen Modellen basieren Rückschlüsse über die interessierende Population. In der klinischen Medizin sind statistische Modelle für die Interpretation von Daten zu Therapieeffekten, zu Diagnose und Prognosestrategien sowie zu Daten in der Versorgungsforschung wichtig. Oft werden statistische Modelle als handhabbare Formen der Wirklichkeit aufgefasst.

Beispielsweise sagen statistische Modelle ein individuelles Schlaganfallrisiko vorher [1] und quantifizieren für bestimmte Patientenkonstellationen die Wahrscheinlichkeit, einen Schlaganfall in einem bestimmten Zeitintervall zu erleiden. Abb. 6 in Rothwell et al. [2] stellt die Ergebnisse aus einem Cox-Modell [3] dar, das aufgrund von 4 Patienteneigenschaften (Ausmaß Stenose, Geschlecht, letztes Ereignis, Zeit seit letztem Ereignis) das 5-Jahres-Risiko für einen ipsilateralen Schlaganfall quantifiziert. Das Bild gibt behandelnden Ärzten Informationen zu den von ihnen betreuten Patienten und repräsentiert klinische Wirklichkeit.

» Es gibt ein weites Spektrum von Modellen für die Analysen der klinischen und der Versorgungsdaten

Kuehnle et al. [4] untersuchen den Zusammenhang von Behandlungsvolumen auf das Schlaganfall- und Mortalitätsrisiko vor Entlassung aus dem Krankenhaus unter Carotisendarterektomie (CEA) bzw. Carotis-Stenting (CAS). Ein Mehrebenen-Poisson-Regressionsmodell [5] untersucht, ob gewisse Volumen-kategorien mit der Häufigkeit von Tod und/oder Schlaganfall in Beziehung stehen. Dabei werden neben bekannten Risikoparametern der Patienten und Institutionen auch unbeobachtete Faktoren mittels Zufallsfaktoren in die Analyse integriert. Die aus dem Modell abgeleiteten Informationen können Einfluss auf Entscheidungen zu Mindestmengen haben.

Ein weites Spektrum von Modellen steht für die Analysen der klinischen Daten und der Versorgungsdaten zur Verfügung. Schaut man in die Literatur, so verwenden Veröffentlichungen oft wenige Modelltypen, die sich in der Praxis durchgesetzt haben. Sind das die wirklich guten Modelle oder ist diese Wahl dadurch bestimmt, dass es das einzig wahre Modell zu einer Fragestellung nicht gibt? Wenn es die eindeutige Modellwahl nicht gibt (bei Rothwell et al. [2] hätte man auch andere Prädiktionsmodelle verwenden können, etwa ein Accelerated-Failure-Time-(AFT-)Modell [6]; Kuehnle et al. [4] hätten auch ein Negativ-Binomialmodell [7] anstelle des Poisson-Modells benutzen können), so stellt sich die Frage, wie findet man ein gutes Modell? Gibt es überhaupt *das gute Modell*? Was ist, wenn es *das gute Modell* nicht gibt und letztendlich ein Spektrum möglicher Modelle für eine angemessene Analyse zur Verfügung steht?

Die Werkzeugkiste der statistischen Modelle

Dieser Abschnitt versucht einen Überblick über verfügbare Regressionsmodelle zu geben. Das Urmodell aller statistischen Modelle ist die *lineare Regression*. Sie untersucht, wie sich der Mittelwert normalverteilter Beobachtungen durch die Ausprägung von Einflussfaktoren bestimmen lässt. Mit der Kenntnis der Streuung kann dann in verschiedenen Gruppierungen der Einflussfaktoren die Verteilung der Beobachtungen bestimmt werden. Die *logistische Regression* [8] modelliert den Einfluss von Faktoren (X) auf die Wahrscheinlichkeit eines binären Ergebnisses ($Y = 0, 1$): $P(Y =$

Hier steht eine Anzeige.

1|X). Sie ist für klinische Anwendungen eines der wichtigsten Modellierungsinstrumente. Die *Poisson-Regression* [5] formalisiert die Häufigkeit von Ereignissen. Sie wird für die Berechnungen von Inzidenzen eingesetzt. Formal müssen beim Einsatz der Poisson-Regression die Mittelwerte von Gruppen gleich der in der Gruppe vorliegenden Varianz sein. Das ist eine wichtige Annahme, die nicht immer erfüllt ist. Oft liegt *Over-Dispersion* in den Daten vor, bei der die entsprechenden Varianzen größer als der Mittelwert sind. Im Falle von *Over-Dispersion* kommt es zur Verwendung einer *Negativ-Binomialregression* [7]. Zeitdauer und ihre Beeinflussung durch Faktoren werden in der Regel mittels *Cox-Modell* oder *proportionalen Hazards-Modells* [6] modelliert. Auch hier liegen Voraussetzungen vor, die nicht immer automatisch erfüllt sind. *Alternative Überlebenszeitmodelle* sind ebenfalls verfügbar [9].

» Die Regressionsdiagnostik prüft die Passgenauigkeit von Modellen an den vorliegenden Datensatz

Es gibt nun verschiedene Techniken, um die Passgenauigkeit von Modellen an den vorliegenden Datensatz zu prüfen. Dies nennt man *Regressionsdiagnostik* [10]. Weiterhin stellt sich die Frage, ob diese Techniken in Fragestellungen mit wenigen oder mit sehr vielen (hochdimensionalen) Einflussgrößen zum Einsatz gebracht werden. Im Fall sehr vieler Einflussgrößen werden *Regularisierungsverfahren* (penalisierte Regression) eingesetzt [11]. Werden Regressionsverfahren zur Erstellung von Prädiktionsregeln verwendet, so können durch *Resampling-Verfahren* und *Kreuzvalidierung* erste Validierungen vorgenommen werden [11].

Weiterhin stellt sich die Frage nach der Kodierung der Einflussgrößen als *kategoriale Variablen* oder als *metrische Variablen* mit speziellen funktionalen Formen (linear, Spline oder Polynom; [10]).

Modellvalidierung und *Modelldiagnostik* sind Gebiete der methodischen Statistik, die sich mit Fragen nach Kriteri-

en und Prüfverfahren beschäftigen, um Modelle hinsichtlich ihrer Passgenauigkeit zu den Daten zu bewerten. Auch hier steht ein großer und spezifizierter Methodenapparat zur Verfügung. Empirische Ergebnisse zeigen, dass die vorhandenen Verfahren es in der Regel nicht erlauben, eindeutig ein optimales Modell aus einer Reihe von konkurrierenden Optionen für die Analyse eines Datensatzes zu wählen.

Zusammenfassend zeigt sich eine sehr reichhaltige Trickkiste statistischer Modelle, zu der es wenig zuverlässige Anleitungen hinsichtlich der Verwendung der verfügbaren Instrumente gibt. Es existiert kein Kanon der perfekten Modelle, aus dem eindeutig für ein bestehendes Projekt der richtige Ansatz gewählt werden kann. Dieses offensichtliche Problem hat massive Konsequenzen für die Praxis: (1) Man beschränkt sich auf das wenige Bekannte, (2) man brilliert als Statistiker und wählt spezielle exotische Modelle, (3) man beruft sich auf fragwürdige Standards oder (4) klammert sich an ein Ritual. Die STRATOS-Initiative (STRengthening Analytical Thinking for Observational Studies) versucht hier Richtlinien zu entwickeln [12] und den Wirrwarr zur Modellwahl zu lichten.

Nutzung statistischer Modelle

Bei der Analyse von *randomisierten kontrollierten Studien* (RCTs) erlauben statistische Modelle die Adjustierung hinsichtlich Einflussgrößen und reduzieren damit die Variabilität in der Studienpopulation. Dies erhöht die Power der Studie und erleichtert das Erkennen von Effekten. Modelle erlauben weiterhin eine Beantwortung von inhaltlich präzisen Fragen zum Therapieeffekt (als nur die Frage: Gibt es einen signifikanten Therapieeffekt?); etwa hinsichtlich prädiktiver Marker (Interaktion zwischen Biomarker und Therapie). Modelle erlauben die Berechnung von Konfidenzintervallen zu relevanten Parametern. Damit liefern sie wichtige Informationen für künftige Studienplanungen und wesentliche Beiträge zur Modellbildung im Health Technology Assessment (HTA) bzw. zu gesundheitsökonomischen Überlegungen. Die Verwendung von Modellen für lon-

gitudinale Daten in RCTs führt zu einem effektiveren Erfassen von Patienteninformationen und erleichtert die Behandlung fehlender Werte durch den Verlust eines Patienten in der Studie.

» Register und klinische Kohorten sind Informationsquellen für Prädiktions- oder Klassifikationsmodelle

Register und *klinische Kohorten* dienen vielfältigen Zwecken. Formal handelt es sich um nichtrandomisierte Beobachtungsstudien. Ihre Verwendung zum Studium von Therapieeffekten oder der Interaktion von Biomarkern und Therapien muss immer mit dem Einfluss von Störgrößen umgehen. Hier ist ein weites Spektrum von Verfahren entwickelt worden, die versuchen Verzerrung in der Schätzung eines Therapieeffektes durch Störgrößen zu eliminieren oder zu minimieren. Ob für das relevante Projekt klassische Adjustierung durch multivariable Modelle ausreicht, um Imbalances zu kontrollieren, oder Methoden des Propensity-Scores verwendet werden, obliegt der Entscheidung des Forschers. Register und klinische Kohorten sind weiterhin relevante Informationsquellen für Prädiktionsmodelle (Prognose) oder Klassifikationsmodelle (Diagnose). Somit sind Fragen zur Validierung von Biomarkern ein zentrales Problem. Hierzu gibt es ein reichhaltiges Methodenspektrum. Neben klassischen statistischen Modellen kommen auch vermehrt Modelle aus dem maschinellen Lernen zum Einsatz (Random Forests, Vector Support Machine, Deep Learning ...; [11]). Eine komplexe Registerstudie aus der Anästhesie zur Analyse einer Interaktion zwischen Geschlecht und Haloperidol ist in [13] zusammen mit den Versuchen durch statistische Modellierung, Bias zu beherrschen, dargestellt.

Versorgungsforschung verwendet in der Regel Modelle, die die hierarchischen Clusterstrukturen der Versorgungsrealität widerspiegeln: Patienten sind in Praxen oder Kliniken eingewiesen, räumliche Strukturen als Einflussfaktor auf Versorgungsleistungen etc.

Bei Analysen basierend auf Routinedaten fehlt oft die präzise Charakterisierung der einbezogenen Patienten und Institutionen. Zufallseffekte erlauben die Berücksichtigung solcher fehlenden Informationen. Für Fragestellungen der Versorgungsforschung haben sich deshalb die Zufallseffektmodelle oder auch Mehrebenenmodelle [14, 15] als wichtiges statistisches Instrumentarium etabliert. Sie können zur Analyse randomisierter Interventionsdaten (clusterrandomisierte Studien; [16]) oder zur Analyse von Beobachtungsdaten [4] verwendet werden. Fragen zum institutionellen Ranking, die oft im Rahmen der Qualitätssicherung diskutiert werden, bedienen sich der gleichen Klasse von Modellen [17].

Das Silberzahn-Uhlmann-Phänomen

Die statistische Analyse gegebener Daten zielt auf *robuste* und *präzise* Ergebnisse. *Präzision* ist in der Regel an der Länge der Konfidenzintervalle zu Schätzern interessierender Parameter erkennbar. Weniger offensichtlich ist, ob diese Ergebnisse *robust* sind. Studienergebnisse sind *robust*, wenn sie nicht von vielfältigen Beschränkungen aufgrund vorliegender Daten, implizit gemachter Annahmen und der durchgeführten Analyse abhängen. *Robuste Studienergebnisse* ändern sich im Wesentlichen nicht, wenn für deren Herleitung angemessene alternative Annahmen und Analysestrategien verwendet werden.

Raphael Silberzahn und Eric L. Uhlmann [18] berichten von einem Experiment bei dem 29 Analyseteams an einem Datensatz der Frage nachgehen, ob bei dunkelhäutigen Fußballspielern das Risiko für einen Platzverweis höher ist als bei hellhäutigen Spielern. Jede der 29 alternativen Analyseansätze war gut begründet und die gefundenen relativen Risiken variierten von einer leicht erhöhten Tendenz bis hin zu einem starken Trend zum Nachteil dunkelhäutiger Spieler. Wäre die zu dieser Fragestellung veröffentlichte und einem Peer-Review unterzogene Arbeit eine dieser 29 Analysen gewesen, so hätte das Ergebnis zwischen keinem bis hin zu großem Einfluss der

Gefäßchirurgie 2017 · 22:412–419 <https://doi.org/10.1007/s00772-017-0306-3>
© Der/die Autor(en) 2017. Dieser Artikel ist eine Open-Access-Publikation.

U. Mansmann

Statistische Modelle in der Gefäßchirurgie

Zusammenfassung

Hintergrund. Die statistische Modellwahl mit dazugehöriger Analysestrategie bestimmt in einem Forschungsprojekt dessen quantitatives Ergebnis. Es existiert in der Regel kein Goldstandard zur Wahl der relevanten Modellierungsstrategie. Der Statistiker kann in Abstimmung mit den weiteren Teammitgliedern im Projekt zwischen einer Vielzahl alternativer Modelle wählen. Folglich kämen verschiedene Teams beim gleichen Datensatz zu verschiedenen Ergebnissen.

Fragestellung. Wie können die für die Wahl von Modell und Analysestrategie gemachten Entscheidungen begründet und transparent dargestellt werden? Wie können diese Entscheidungen unparteiisch gefällt werden? Wie kann es zum Konsens zwischen verschiedenen Forschungsteams in der Beantwortung einer Fragestellung kommen?

Material und Methoden. Diskussion von Beispielen.

Ergebnisse. Die Wahl einer statistischen Modellierung und Analyse sollte auf folgende Konzepte eingehen: Transparenz, Konsens, Unparteilichkeit, Bezug auf die beobachtete Realität, Wahrung multipler Perspektiven, Offenlegung der Kontextabhängigkeit sowie Robustheit der Ergebnisse.

Schlussfolgerung. Die Komplexität der statistischen Modellierung erlaubt in der Regel nur selten Objektivität bei der Durchführung einer Analyse. Zur Darstellung der entsprechenden Entscheidungsmöglichkeiten werden neue Strategien und Formate benötigt.

Schlüsselwörter

Statistische Modelle · Studiendesgin · Wahrscheinlichkeit · Konsens · Objektivität

Statistical modelling in vascular surgery-related clinical research

Abstract

Background. The choice of a statistical model for the analysis of specific data determines the results of a study. Unfortunately, a true perfect model does not exist. There are various options to shape model choice and analysis strategy with possibly very divergent quantitative study results. It is probable that different research teams will come up with different results from the same data set.

Objectives. How is it possible to justify and to transparently report the process of model building together with the specification of the analysis strategy? How is it possible to implement impartiality into this development process? How can consensus building be improved between different research teams regarding the same research question?

Methods. Presenting and discussing examples.

Results. Principles of good science in selecting a statistical model require compliance with the following concepts: transparency, consensus, impartiality and correspondence to observable reality, awareness of multiple perspectives and context dependence, and investigation of stability.

Conclusions. The complexity of statistical model choice rarely allows performing an analysis objectively. New forms of reporting the decisions made behind an analysis are needed.

Keywords

Models, statistical · Research design · Probability · Consensus · Objectivity

Rasse des Fußballspielers liegen können. Die Variabilität der Ergebnisse verdeutlicht den Einfluss subjektiver Entscheidungen zur Analysestrategie innerhalb der 29 Teams.

Für viele Wissenschaftler ist das weite Spektrum der aus einem Datensatz abgeleiteten möglichen Effekte verwirrend. Sie fragen, ob überhaupt das richtige Modell für die Analyse verwendet wurde? Sie

glauben an ein richtiges, objektives Vorgehen. Aber vielleicht ist es falsch, eine einzelne Analyse zu ernst zu nehmen.

Das von Silberzahn und Uhlmann berichtete Phänomen ist allgegenwärtig. Es ist kaum sichtbar, weil nur selten mehrere Teams simultan, unabhängig voneinander einen gegebenen Datensatz zu einer gegebenen Frage auswerten. Es zeigt sich jedoch bei systematischen Reviews ran-

domisierter kontrollierter Studien. Die dort beobachtete Variabilität der Therapieeffekte hat neben den subjektiven Entscheidungen bei der statistischen Analyse noch weitere Komponenten: Unterschiede in den Protokollen, Unterschiede in der Umsetzung der Studie. Die Robustheit der Aussage zum Therapieeffekt wird in solchen systematischen Reviews durch die Analyse der Verzerrungsfälligkeit und durch den Forest-Plot [19] der Studienergebnisse sichtbar. Weiterhin gibt es mannigfaltige Versionen von Prädiktionsalgorithmen zu spezifischen Fragestellungen. Bedeutet dies, dass alle in der Literatur aufgeführten Algorithmen falsch sind, oder führen diese trotz verschiedener Einflussgrößen zu kongruenten Prädiktionen? Bei Studien aus dem Versorgungssetting (Mindestmengen, Qualitätsbewertung) ist in der Regel nichts über die Robustheit der Ergebnisse gegenüber der Wahl der statistischen Strategie bekannt. Zu welcher Variabilität von Ergebnissen kann das Silberzahn-Uhlmann-Phänomen hier führen?

Gibt es das objektiv richtige Modell?

Die Existenz des objektiv richtigen Modells und der objektiv richtigen Analyse zur Beschreibung empirischer Daten wird von Statistikern angezweifelt. Georg Box sagt: „Essentially, all models are wrong, but some are useful“ [20, S. 424]. Es besteht somit grundsätzlich Unsicherheit über das richtige (passende, nützliche) Modell.

Modelle dürfen nicht mit der Wirklichkeit verwechselt werden. Gesucht werden Modelle, die als alternative Analysen zu einer Frage an einem Datensatz passen und einen robusten und reliablen Blick auf die hinter den Daten stehende Wirklichkeit erlauben. Ergebnisse, die robust bleiben gegenüber verschiedenen passenden Modellen, werden möglicherweise als objektiv existierend anerkannt. Wie kann ein Forscherteam ausreichend viele optionale Modelle entwickeln und einsetzen? Ist zum Verständnis einer Forschungsfrage die von Silberzahn und Uhlmann [18] beschriebene Crowd-Analyse notwendig? Welchen Entschei-

dungsprozessen folgt ein Forscherteam um die eigenen Analysestrategien zu finden? Was kann die Modellwahl beeinflussen und zu einer begründeten Vielfalt von Modellen führen?

Gelman und Hennig [21] geben hierzu eine kurze Liste von Entscheidungen, die bei der statistischen Modellierung und Festlegung der Analysestrategie aus dem weiten Spektrum von Möglichkeiten zu fällen sind:

1. Empfehlungen zur Wahl von Variablen, nach denen im Modell adjustiert werden soll,
2. Umgang mit Ausreißern in den Daten: Welche sind wirkliche Ausreißer? Welche ergaben sich aufgrund von Messfehlern? Welche Ausreißer entstehen dadurch, dass die Messung korrekt ist, das Individuum aber nicht in die interessierende Population passt?
3. Umgang mit Messproblemen und Störgrößen (Confounding),
4. Definition von Effektmaßen (absolut, relativ?),
5. Gewinnung des Datensatzes aus der Grundpopulation (Fehlwerte),
6. Festlegen von Glattheitsbedingungen und Einschränkungen auf niedrigdimensionale Situationen,
7. Festlegung der funktionellen Form der Einflussgrößen (kategorial, stetig, linear, nichtlinear),
8. gibt es numerische Informationen zu Modellparameter (Streuung, bekannte Risiken)?

Anstelle der Suche nach dem objektiv richtigen Modell bietet sich somit für ein Forscherteam eine sorgfältige Dokumentation und Begründung der durchgeführten Analyseschritte an. Der Platz für eine solche Dokumentation ist der statistische Analyseplan (SAP). In der Regel werden SAPs selten veröffentlicht. Sie sind oft eine trockene und technische Lektüre. Ihre Offenlegung trägt aber wesentlich zur wissenschaftlichen Qualität eines Projektes bei. Ein SAP und die Studiendaten sind die Grundvoraussetzung für die Reproduzierbarkeit der Studienergebnisse durch externe Gruppen [22].

Wie soll man mit Modellunsicherheit umgehen?

Transparenz ist notwendig, um alle impliziten und expliziten Entscheidungen darzulegen und zu rechtfertigen, die zur statistischen Analyse geführt haben. Die Angabe externer und nachprüfbarer Quellen macht das subjektive Vorgehen nachvollziehbar. Die Dokumentation gibt Einblick in die Verwendung externer Information und wie daraus Modellannahmen entstanden sind. Sie dokumentiert die Entscheidungen, die der durchgeführten Analyse und den daraus gezogenen Schlüssen zugrunde liegen. Diese Transparenz umfasst weiterhin die impliziten und expliziten Annahmen über die Entstehung der Daten für die folgende Analyse. Hierzu zählen beispielsweise die Annahmen, ob fehlende Werte zufällig entstanden sind („missing at random“, MAR; [23]), ob Einflussgrößen fehlerfrei oder mit Messfehler erhoben wurden [24]. Transparenz, die die Hintergründe und Einflüsse einer Analyse offenlegt, trägt zur *Konsensbildung* bei.

Wissenschaftlicher *Konsens* basiert auf Begründungen, klaren Argumenten und Offenlegung der Motive und der Darlegung, wie dies mit dem bestehenden Sachwissen in Beziehung steht. *Konsens* ist ein Gruppenprozess. Ein Analyst sollte deshalb seine Arbeitsweise so klar darstellen, dass mit seiner Darstellung die *Konsensbildung* gestärkt wird. *Konsensbildung* bedingt weiterhin, dass die Analyse unparteiisch durchgeführt wurde, dass sie konkurrierenden Perspektiven Rechnung trägt, klar die Bevorzugung von Lieblingshypothesen vermeidet und offen für Kritik ist.

Silberzahn und Uhlmann [18] implizieren zwei weitere wichtige Aspekte: (1) *Aufmerksamkeit gegenüber multiplen Perspektiven*. Breiman nennt dies das Rashomon-Phänomen [25]. In dem japanischen Roman *Rashomon* beobachten vier Personen ein Verbrechen. Vor Gericht berichten sie die gleichen Fakten aber als vier völlig verschiedene Geschichten. (2) *Offenlegung des Kontextes*, in dem der Statistiker seine Analyse entwickelt hat: spezifischer individueller Hintergrund und Forschungsinteressen

Tab. 1 Die Tugenden einer guten statistischen Analyse nach Gelman und Hennig [6]

Tugend	Aspekte
Transparenz	Klare und eindeutige Konzeptdefinitionen
	Offene Planungsunterlagen
	Analyse folgt einem klaren Protokoll
	Offenlegung der Gründe, Prozeduren, Annahmen und Beschränkungen
Konsens	Referenz auf bestehendes relevantes Wissen
	Referenz auf bestehende verwandte Arbeiten
	Befolgen allgemeiner Regeln wenn möglich und passend
	Offenlegung der Beweggründe für Konsensbildung und Vereinheitlichung
Unparteilichkeit	Sorgfältige Diskussion relevanter und potenziell konkurrierender Theorien und Betrachtungsweisen
	Sorgfältiger Umgang mit Störgrößen (Confoundern)
	Offenheit für Kritik und Ideenaustausch
Bezug auf die beobachtete Realität	Wie sind Konzepte und Modell mit den Daten verbunden?
	Welche Bedingungen sind zu erfüllen, um die Ergebnisse zu reproduzieren, testen oder falsifizieren zu können?
Wahrung multipler Perspektiven	Formulierung und Diskussion alternativer Modellansätze und Erklärungsmöglichkeiten, deren Umsetzung in Sensitivitätsanalysen
Offenlegung der Kontextabhängigkeit	Darstellung der Abhängigkeit von Kontext und Zielen der Arbeit
	Darlegung der Position, der Ziele, der Erfahrung des Analysierenden
	Darstellung subjektiver Gesichtspunkte
Robustheit der Ergebnisse	Welche Konsequenzen haben mögliche Entscheidungen während der Analyse auf die Interpretation der Ergebnisse?
	Wie variabel sind die Ergebnisse? Lassen sie sich mit neuen Daten validieren und reproduzieren?

sowie Perspektiven des Projektes. Es ist von Interesse zu verstehen, wie der Kontext und die Ziele eines Projektes mit den Entscheidungen zusammengehen, die die Analysestrategie festlegen.

Die Tugenden von Gelman und Hennig

Gelman und Hennig [21] sehen in der Objektivierung statistischer Modellierung ein grundsätzliches wissenschaftliches Problem. Objektive statistische Modellbildung ist nur in Ausnahmefällen und in sehr einfachen Settings möglich. Praktisch führt dieser Objektivierungsanspruch der statistischen Modellbildung zum Verschweigen wichtiger Informationen, zu fehlender Nachvollziehbarkeit wichtiger Ergebnisse und zur Beschränkung der wissenschaftlichen Kreativität. Gelman und Hennig empfehlen die Dichotomie zwischen *objektiv* und *subjektiv* durch einen Katalog von Tugenden zu ersetzen, die in **Tab. 1** formuliert sind.

Ausblick und Diskussion

Die Darstellung der Modellbildung in den Beiträgen zur klinischen Forschung ist extrem begrenzt. In Fachzeitschriften wird weder der Platz in entsprechenden Veröffentlichungen zur Verfügung gestellt noch besteht ein tieferes Interesse an diesem Thema bei den medizinischen Kollegen. Teilweise wird dies damit begründet, dass es objektiv das gute und richtige Modell gibt und der Projektstatistiker, dieses auch verantwortungsvoll umgesetzt hat. Dennoch sind viele Ergebnisse nicht eindeutig reproduzierbar, welche Rolle dem Statistiker dabei zukommt, wird im Moment kaum diskutiert.

Silberzahn und Uhlmann [18] haben in ihrem kleinen Experiment gezeigt, dass selbst bei gleichem Datensatz und einer gleichen sehr präzise gestellten Fragestellung (präzise Quantifizierung eines relativen Risikos) sorgfältig arbeitende Statistiker zu einem weiten Spektrum von Antworten kommen. Es gibt also prinzipiell einen starken Effekt

der Modell- und Analysewahl auf das Ergebnis eines Projektes. Dieses Phänomen ist sowohl bei sorgfältig geplanten randomisierten kontrollierten Studien zu erwarten als auch bei der Analyse von Beobachtungsdaten.

Vance Berger hatte 2008 dieses Problem für RCTs adressiert [26] und vorgeschlagen, einen Kanon guter Auswertungsverfahren (die den State of the Art repräsentieren) aufzustellen, auf den die statistische Community zurückgreifen kann. Ist ein Statistiker bei der Analyse seines Problems davon überzeugt, dass die im Kanon vorgegebene Musterlösung unpassend ist, so kann er in entsprechenden Sensitivitätsanalysen davon abweichen, wenn er diese Abweichung wissenschaftlich begründen kann. Im Moment ist ein solcher Methodenkanon noch nicht konsentiert. Die STRATOS-Bewegung versucht dies für die Analyse von Beobachtungsstudien [12]. Zur Analyse von RCTs ist ein solcher Kanon ebenfalls nicht verfügbar. Vielmehr existiert eine *ICH Guideline on Statistical Principles for Clinical Trials (ICH E9)* [27], die Aussagen über Konzepte der Auswertung macht, aber keine Empfehlung über die Verwendung spezifischer Methoden und Modelle ausspricht. Somit ist der von Vance Berger vor 10 Jahren geforderte Kanon als Ausgangspunkt für statistische Modellbildung und Formulierung statistischer Analysestrategien noch nicht verwirklicht.

» Die Verantwortung zur Erstellung eines statistischen Analyseplans tragen die Projektwissenschaftler

Die volle Verantwortung zur Erstellung eines statistischen Analyseplans (SAP) liegt somit in der Hand der Projektwissenschaftler. Das Dokument kann neben der trockenen statistischen Programmierung aber auch Kommentare enthalten, die die Vorgehensweise auch für die nichtstatistikbewanderten Kollegen interessant macht. Beim Verfassen eines Analyseplans können die Tugenden von Gelman und Hennig [21] problemlos berücksichtigt werden. Es gibt weiterhin

Plattformen, auf denen ein SAP zur Diskussion und zum Review eingestellt werden kann. Als Beispiel kann [28] dienen. Nach Einstellen des Dokumentes fand ein internationaler Review durch vier interessierte Kollegen statt. Nach dem Review und dessen Einarbeitung in das Dokument wurde der SAP auf PubMed verfügbar gemacht. Dieser Prozess hat das Dokument in größere Übereinkunft mit den Tugenden von Gelman und Hennig gebracht.

Die Zukunft wird eine Zunahme an veröffentlichten SAPs bringen. Die Reproduzierbarkeitskrise in der Biomedizin [29] wie auch in anderen Wissenschaftsbereichen (Psychologie und Wirtschaft) wird diese Entwicklung erzwingen. Was diesbezüglich für die vorklinische Forschung gilt, ist auch in der klinischen Forschung relevant. Ein gewisser Beitrag muss hier aus der Statistik kommen: die Entwicklung von Instrumenten zur umfassenden Offenlegung der Konzeptionierung und Durchführung einer statistischen Analyse und der dabei gewählten statistischen Modelle. Damit lassen sich wissenschaftliche Kreativität und methodische Strenge produktiv verbinden.

Fazit für die Praxis

- Ob RCT oder Beobachtungsstudie, es stehen eine Vielzahl von Modellierungsansätze zur Verfügung, um die Forschungsfrage statistisch zu bearbeiten.
- Die Modellwahl bestimmt das quantitative Ergebnis, eine objektive Modellwahl ist selten möglich.
- Dies eröffnet die Möglichkeit eines breiten Ergebnisspektrums für die Analyse.
- Analysestrategie und Modellwahl müssen unparteiisch begründet und transparent dargestellt sein;
- Modellierung und Analysestrategie müssen eine klar nachvollziehbare Verbindung zum beobachteten Sachverhalt besitzen.
- Die Analyse muss multiple Perspektiven und Kontextabhängigkeit berücksichtigen.
- Die Analyse muss versuchen robust gegenüber Annahmen zu sein.

- Zur Reproduzierbarkeit von Studienergebnissen wird es notwendig werden, ausführliche Analysepläne zu veröffentlichen. Wissenschaftliche Zeitschriften sollten diese in den zum Artikel gehörenden elektronischen Supplements zur Veröffentlichung anbieten.

Korrespondenzadresse

Prof. Dr. U. Mansmann
Institut für medizinische
Informationsverarbeitung,
Biometrie und Epidemiologie,
Ludwig Maximilians
Universität München
Marchioninistraße 15,
81377 München, Deutschland
mansmann@
ibe.med.uni-muenchen.de

Einhaltung ethischer Richtlinien

Interessenkonflikt. U. Mansmann gibt an, dass kein Interessenkonflikt besteht.

Dieser Beitrag beinhaltet keine von den Autoren durchgeführten Studien an Menschen oder Tieren.

Open Access Dieser Artikel wird unter der Creative Commons Namensnennung 4.0 International Lizenz (<http://creativecommons.org/licenses/by/4.0/deed.de>) veröffentlicht, welche die Nutzung, Vervielfältigung, Bearbeitung, Verbreitung und Wiedergabe in jeglichem Medium und Format erlaubt, sofern Sie den/die ursprünglichen Autor(en) und die Quelle ordnungsgemäß nennen, einen Link zur Creative Commons Lizenz beifügen und angeben, ob Änderungen vorgenommen wurden.

Literatur

1. Rothwell PM, Warlow CP, European Carotid Surgery Trialists' Collaborative Group (1999) Prediction of benefit from carotid endarterectomy in individual patients: a risk modeling study. *Lancet* 353:2105–2110
2. Rothwell PM, Mehta Z, Howard SC, Gutnikov SA, Warlow CP (2005) From subgroups to individuals: general principles and the example of carotid endarterectomy. *Lancet* 365:256–265
3. Cox D (1972) Regression models and life tables. *J R Stat Soc Ser B* 34:187–220
4. Kuehn A, Tsantilas P, Knappich C, Schmid S, König T, Breikreuz T, Zimmermann A, Mansmann U, Eckstein HH (2016) Significant association of annual hospital volume with the risk of in-hospital stroke or death following carotid endarterectomy but likely not after carotid stenting: secondary data analysis of the statutory German carotid quality assurance database. *Circ Cardiovasc Interv* 9(11):e4171
5. Jones AM et al (2013) „Models for count data“. *Applied health economics*. Routledge, London, 5295–341
6. Collett D (2003) *Modelling survival data in medical research*, 2. Aufl. CRC press, Boca Raton
7. Cameron AC, Trivedi PK (1998) *Regression analysis of count data*. Cambridge University Press, Cambridge
8. Collett D (2002) *Modelling binary data*. Chapman & Hall, CRC, Boca Raton
9. Klein JP, Moeschberger ML (2003) *Survival analysis – techniques for censored and truncated data*. Springer, New York
10. Fahrmeir L, Kneib T, Lang S, Marx B (2013) *Regression – Models, Methods and Applications*. Springer, Berlin Heidelberg
11. Hastie T, Tibshirani R, Friedman J (2009) *The elements of statistical learning*. Springer, Berlin, Heidelberg
12. <http://www.stratos-initiative.org/>. Zugriffen: 30.06.2017
13. Brettner F, Janitza S, Prüll K, Weninger E, Mansmann U, Küchenhoff H, Jovanovic A, Pollwein B, Chappell D, Zwissler B, von Dossow V (2016) Gender-specific differences in low-dose Haloperidol response for prevention of postoperative nausea and vomiting: a register-based cohort study. *PLOS ONE* 11(1):e146746
14. Goldstein H (2011) *Multilevel statistical models*, 4. Aufl. Wiley, Chichester
15. Spiegelhalter DJ, Myles JP, Jones DR, Abrams KR (1999) *Methods in health service research. An introduction to bayesian methods in health technology assessment*. *BMJ* 319(7208):508–512
16. Campbell MJ, Donner A, Klar N (2007) Developments in cluster randomized trials and statistics in medicine. *Stat Med* 26:2–19
17. Henschel V, Engel J, Hölzel D, Mansmann U (2009) A semiparametric Bayesian proportional hazards model for interval censored data with frailty effects. *BMC Med Res Methodol* 9:9
18. Silberzahn R, Uhlmann EL (2015) Many hands make tight work. *Nature* 526:189–191
19. Lewis S, Clarke M (2001) Forest plots: trying to see the wood and the trees. *BMJ* 322(7300):1479–1480
20. Box GEP, Draper NR (1987) *Empirical model building and response surfaces*. John Wiley & Sons, New York
21. Gelman A, Hennig G (2017) Beyond subjective and objective in statistics. *J R Stat Soc Ser A* 180(4):1–31
22. Sandercock P, Lindley R, Wardlaw J, Whiteley W (2012) Statistical analysis plan for the third International Stroke Trial (IST-3); part of a „thread“ of reports of the trial. *Int J Stroke* 7:186–187
23. Sterne JA, White IR, Carlin JB, Spratt M, Royston P, Kenward MG, Wood AM, Carpenter JR (2009) Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *BMJ* 338:b2393
24. Hutcheon JA, Chiolero A, Hanley JA (2010) Random measurement error and regression dilution bias. *BMJ* 340:c2289
25. Breiman L (2001) Statistical modeling: the two cultures (with comments and a rejoinder by the author). *Stat Sci* 16(3):199–231
26. Berger VW, Matthews JR, Grosch EN (2008) On improving research methodology in clinical trials. *Stat Methods Med Res* 17(3):231–242
27. <http://www.ich.org/products/guidelines/efficacy/efficacy-single/article/statistical-principles-for-clinical-trials.html>. Zugriffen: 30.6.2017

28. Dallaku K, Shakur H, Edwards P, Beaumont D, Roberts I, Huque S, Delius M, Mansmann U (2016) Statistical analysis plan for the WOMAN-ETAPlaT study: Effect of tranexamic acid on platelet function and thrombin generation. Wellcome Open Res 1:30
29. Freedman LP, Cockburn IM, Simcoe TS (2015) The economics of reproducibility in preclinical research. Plos Biol 13(6):e1002165

Einfrieren roter Blutkörperchen mit natürlichem „Frostschutz“

Ein internationales Wissenschaftler-Team konnte jetzt nachweisen, dass das natürliche Frostschutzmittel Trehalose durch die Anwesenheit von mineralischen Nanopartikeln (Apatit) in die Blutzellen geschleust werden kann

Rote Blutkörperchen zur Bevorratung werden bei minus 80 Grad Celsius gelagert. Als Gefrierschutz nutzt man seit den 1970er Jahren Lösungsmittel wie Glycerin. Da dieses giftig ist, muss es vor der Transfusion mittels Dialyse aufwändig aus dem Blut entfernt werden. Im Gegensatz dazu könnten natürliche Frostschutzmittel wie Trehalose – ein Disaccharid, das in einigen Pflanzen und Pilzen vorkommt – bei der Kryokonservierung genutzt werden, ohne dass eine anschließende Dialyse notwendig wäre. Allerdings hat sich gezeigt, dass auf diese Weise beim Auftauen zu viele rote Blutzellen absterben.

Nanopartikel als Hilfsmittel

Um zu erforschen, wie die Kryokonservierung mittels Trehalose dennoch gelingen könnte, hat eine Forschergruppe vom CIRIMAT-Institut der Universität Toulouse Experimente mit Apatit-Nanopartikeln durchgeführt (Durchmesser circa 40 Nanometer). Diese mineralischen Partikel konnten die Wissenschaftler kostengünstig selber herstellen. Sie bestehen aus denselben Bestandteilen wie Knochen oder Zähne, sind also biokompatibel. In Zusammenarbeit mit Materialwissenschaftlern aus Slowenien testeten sie die Kryokonservierung der roten Blutzellen mittels Trehalose in Anwesenheit von Nanopartikeln. In diesen Versuchen überstanden mehr als 90 % der roten Blutkörperchen den Auftauprozess unbeschadet; damit wäre dieses Verfahren für klinische Anwendungen praktikabel. Die Forscher schlussfolgerten, dass die Apatit-Nanopartikel die Aufnahme von Trehalose durch die Zellmembran der roten Blutkörperchen ermöglichen.

Gefrierschutz Trehalose

Welcher Mechanismus dabei wirksam wurde, untersuchte ein Team von theoretischen Physikern der spanischen Universität „Rovira i Virgili“ und von Experimentalphysikern um Dr. Jean-Baptiste Fleury an der Universität des Saarlandes. Die Saarbrücker Forscher entwarfen eine künstliche Zellmembran aus einer Doppellage von Phospholipiden und

untersuchten die Interaktion einzelner Nanopartikel mit der Membran. Mittels Fluoreszenzmikroskopie und elektrophysiologischen Messungen konnten sie ermitteln, mit welcher Geschwindigkeit sich die Nanopartikel an die Membran anhefteten. Diese gilt als Maß für die Adhäsionsenergie und ist die entscheidende Größe für die anschließende Computersimulation der spanischen Wissenschaftler: In dieser konnte die Membran um einzelne Nanopartikel-Teilchen herum dargestellt werden. Es zeigte sich, dass sich in der Umgebung der Nanoteilchen die „Packungsdichte“ der Phospholipid-Moleküle verringerte, wodurch die Membran – ähnlich wie ein Gummiband – lokal gedehnt wurde. Als Folge kann der Gefrierschutz Trehalose die Zellmembran in ausreichend großer Menge durchdringen.

**Quelle: Universität des Saarlandes
www.sfb1027.uni-saarland.de**

**basierend auf: Biomaterials (2017)
140:138-149**