Dr. Jens Dreyhaupt

Instrumente für Power- und Fallzahlberechnungen bei komplexen hierarchischen Studiendesigns in der Versorgungsforschung

Zusammenfassung

Hintergrund: Studien stellen wichtige Instrumente im Erkenntnisgewinn in der Versorgungsforschung dar, wobei oft Studiendesigns mit hierarchischer Datenstruktur und zusätzlichen Kovariaten auftreten (Studiendesigns mit komplexer hierarchischer Datenstruktur). Beispiele sind longitudinale Studien zur Untersuchung von Zeitverläufen (z.B. Körpergewicht) mit Berücksichtigung zusätzlicher Einflussgrößen (z.B. Geschlecht, Alter zu Baseline) oder clusterrandomisierte Studien mit zusätzlicher Berücksichtigung weiterer Clustereigenschaften (z.B. Migrantenanteil, Baseline-Wert der Zielgröße). Werden für solche Designs konfirmatorische Fragestellungen untersucht, ist eine a-priori Abschätzung der Fallzahl bzw. statistischen Power erforderlich, wofür in komplexeren Situationen nur unzureichende oder keine Methoden verfügbar sind.

Methoden: Möglichkeiten der Power- bzw. Fallzahlabschätzung bei komplexen hierarchischen Datenstrukturen wurden über umfangreiche Recherchen eruiert. Die Ergebnisse wurden anschließend auf Verwendbarkeit geprüft.

Ergebnisse: Es wurde eine frei verfügbare Software gefunden, mit welcher Programmcode für Powerabschätzungen via Simulation bei metrischen und binären Zielgrößen für viele gängige Situationen in der Versorgungsforschung erzeugt werden kann. Zur nutzerfreundlichen Anwendung dieser Software wurde eine Anwenderbroschüre erstellt. Für spezielle Studiendesigns wurden Formeln gefundenen und in einem zweiten Bericht, der Formelsammlung, zusammengestellt. Für die Planung und Durchführung von Feasibility- oder Pilotstudien im Kontext komplexer hierarchischer Datenstrukturen werden in einem dritten Bericht Hinweise gegeben. Im Artikel wird die beispielhafte Nutzung von Anwenderbroschüre und Formelsammlung zur Planung einer Studie mit einer metrischen Zielgröße behandelt.

Diskussion/Fazit: Mit den drei Berichten werden Instrumente zur Power- oder Fallzahlabschätzung bei metrischen und binären Zielgrößen für eine Vielzahl konfirmatorischer Studien mit komplexer hierarchischer Datenstruktur bereitgestellt.

Schlüsselwörter

hierarchische Studiendesigns mit Kovariate, Power- und Fallzahlberechnung, Simulation, cluster-randomisierte Studien

Die Versorgungsforschung kann in ein System mit mehreren Ebenen eingeordnet werden und greift auf jeder Ebene auf eine Reihe von Studiendesigns zurück (Hey/Maschewsky-Schneider 2006: Kapitel 6; Pfaff et al. 2011). Allgemein stellen Studien wichtige Instrumente im Erkenntnisgewinn in der Versorgungsforschung dar (z.B. Meyer et al. 2012; Karbach et al. 2012). Hierbei finden immer häufiger Studiendesigns mit mehreren Hierarchieebenen (sogenannte Multilevel-Datenstrukturen) Anwendung. Ein bedeutender Aspekt in der Versorgungsforschung ist die Notwendigkeit der Adjustierung für mögliche wesentliche Confounder, „da versorgungsrelevante Parameter im Mittelpunkt der Analyse stehen" (Hey/Maschewsky-Schneider 2006: 251). Hieraus können sich Studiendesigns mit mehr als einer Hierarchieebene und zusätzlicher Berücksichtigung von Kovariaten ergeben. Diese Designs werden im Folgenden als Studiendesigns mit komplexer hierarchischer Datenstruktur bezeichnet. Beispiele sind longitudinale Studien zur Untersuchung von Zeitverläufen (z.B. Körpergewicht) mit Berücksichtigung zusätzlicher Einflussgrößen (z.B. Geschlecht, Alter zu Baseline) oder clusterrandomisierte Studien mit zusätzlicher Berücksichtigung weiterer Clustereigenschaften (z.B. Migrantenanteil, Baseline-Wert der Zielgröße). Ein konkretes Beispiel einer solchen Studie ist URMEL-ICE, eine cluster-randomisierte Studie zur Untersuchung der Wirksamkeit eines schulbasierten Programms zur Prävention von Übergewicht bei Grundschulkindern (Brandstetter et al. 2012). Neben der Gruppenzugehörigkeit (Interventionsgruppe vs. Kontrollgruppe) wurden auch Baselinewert und Zeitabstand der Messungen in der Analyse berücksichtigt.

>> Eine zentrale Frage ist dabei, ob mit einer Studie eine für die Fragestellung ausreichend große Strichprobe untersucht wurde, da dies eine wesentliche Anforderung an Studiendesign und Studienpopulation darstellt (z.B. Pfaff et al. 2009: 510). Da auch in der Versorgungsforschung konfirmatorische Untersuchungen bei Studiendesigns mit komplexen hierarchischen Datenstrukturen durchgeführt werden, besteht hier die Aufgabe einer adäquaten Fallzahlberechnung bzw. Powerabschätzung. Beispiele hierfür sind Registerstudien (Pfaff et al. 2011), cluster-randomisierte Studien (Eldridge/Kerry 2012) oder Studien, in denen longitudinale Daten entstehen (Fitzmaurice et al. 2011). Bei komplexen hierarchischen Datenstrukturen können Messwerte der Beobachtungseinheiten meist nicht als statistisch unabhängig betrachtet werden. Eine adäquate Powerabschätzung bzw. Fallzahlplanung erfordert daher die Berücksichtigung dieser Abhängigkeiten. Ein Ignorieren kann zur Überschätzung der Power bzw. einer zu kleinen Fallzahl führen (z.B. Eldridge/Kerry 2012; Dreyhaupt et al. 2013a), was aus ethischen und ökonomischen Gründen problematisch ist.

Methoden der Fallzahlplanung bei einfachen Studiendesigns mit komplexer hierarchischer Datenstruktur sind beispielsweise in Donner/Klar 2010; Fitzmaurice et al. 2011; Eldridge/Kerry 2012 und Heo/Leon 2008 zu finden. Diese Methoden sind jedoch bei komplexeren Studiendesigns (wie beispielsweise Multilevel-Datenstrukturen mit mehr als einer Hierarchieebene und Berücksichtigung zusätzlicher Kovariate) nur unzureichend oder nicht anwendbar. Eine Lösung besteht dann in der Anwendung von Simulationsverfahren (Monte-Carlo-Simulation, siehe z.B. Dang et al. 2008; Bell et al. 2010; Reich et al. 2012; Kontopantelis et al. 2014; Meinck/Vandenplas 2012), die auch in der Versorgungsforschung angewendet werden können (z.B. Adams et al. 2003). Mit Simulationsverfahren können neben spezifischen pragmatischen Anforderungen, wie beispielsweise eine im Vorfeld limitierte Anzahl an Clustern (z.B. die begrenzte Verfügbarkeit von Reha-Zentren oder Pflegeheimen mit einer bestimmten Ausstattung), auch unsichere Annahmen in der Phase der Studienplanung (z.B. Clustergröße, Clusteranzahl) vergleichsweise einfach Berücksichtigung finden.

In diesem Artikel werden drei Instrumente vorgestellt, die in der Versorgungsforschung und anderen Gebieten zur Fallzahl- und Powerabschätzung für Studien mit metrischen und binären Zielgrößen mit bis zu drei Hierarchieebenen und Berücksichtigung von Kovariaten dienen können. Beschrieben wird eine Anwenderbro-

schüre zum nutzerfreundlichen Gebrauch einer Spezialsoftware zur Powerabschätzung für eine Vielzahl von Studiendesigns. Weiterhin wird eine Formelsammlung vorgestellt, mit Formeln zur Fallzahl- und Powerabschätzung für einige spezielle komplexe hierarchische Studiendesigns. Abschließend wird auf einen dritten Bericht eingegangen, welcher Informationen zur Fallzahl- und Powerabschätzung im Kontext komplexer hierarchischer Datenstrukturen bei fehlenden Vorinformationen enthält.

Dieser Artikel entstand im Rahmen eines Forschungsprojekts, welches innerhalb des Nachwuchsprogramms des Netzwerks ‚Versorgungsforschung Baden-Württemberg' durchgeführt wurde, das vom baden-württembergischen Ministerium für Wissenschaft, Forschung und Kunst und dem Ministerium für Arbeit und Sozialordnung, Familie, Frauen und Senioren gefördert wird.

Methoden

Zu Beginn erfolgten zunächst umfassende Internet- und Literaturrecherchen zu Möglichkeiten der Power- bzw. Fallzahlabschätzung bei komplexen hierarchischen Datenstrukturen, wobei neben aktuellen Publikationen insbesondere nach verfügbarer Spezialsoftware gesucht wurde. Hierfür wurde neben der Suchmaschine Google auch die Datenbank PubMed einbezogen. Für die Projektziele hat sich die Software MLPowSim (Browne et al. 2009) als geeignet erwiesen. Zu deren nutzerfreundlichen Anwendung wurde im Rahmen des Projekts eine Anwenderbroschüre erstellt. Weiterhin wurde in Publikationen nach „einfach anwendbaren" Formeln für Power- oder Fallzahlabschätzung in komplexen hierarchischen Datenstrukturen gesucht. Diese Formeln wurden in einem zweiten Bericht, der Formelsammlung, zusammengestellt. Da sowohl MLPowSim als auch die Formelsammlung das Vorliegen von Vorinformationen (Literatur, Pilotstudie, ähnliche Projekte) voraussetzen, wurden ein dritter Bericht zur Power- und Fallzahlabschätzung für Situationen ohne verfügbare Vorinformationen im Kontext komplexer hierarchischer Daten erstellt. Diese drei Berichte bieten eine Unterstützung zur Power- und Fallzahlberechnung für komplexe hierarchische Studiendesigns und sollen insbesondere dem Anwender in der Versorgungsforschung eine Hilfestellung zur Studienplanung geben.

Ergebnisse

Anwenderbroschüre

Die Anwenderbroschüre ist ein Bericht, in welchem die Nutzung der Software MLPowSim zur Erzeugung von Quellcode eines Simulationsprogramms sowie die Ergebnisse der Ausführung dieses Programms beschrieben ist. MLPowSim ist eine freie Software, die als Download einer ausführbaren Datei (powerupdate.exe) auf der Internetseite http://www.bristol.ac.uk/cmm/software/mlpowsim/ unter dem Link „MLPowSim executable file (exe, 0.7 mb)" verfügbar ist (Abruf: 29.01.2015). Sie kann zur Powerabschätzung sowohl für metrische als auch binäre Zielgrößen in komplexen hierarchischen Studiendesigns mit bis zu drei Hierarchieebenen verwendet werden. Mit MLPowSim kann für die Statistikpakete R oder MLwiN Quellcode eines Simulationsprogramms zur Powerabschätzung erzeugt werden. Dieser Quellcode muss dann in dem entsprechenden Statistikpaket geöffnet und gestartet werden. Als Resultat des Programmablaufs wird eine Powerabschätzung für Settings eines vorher in MLPowSim spezifizierten Studiendesigns erhalten. Die Anwendung von MLPowSim wurde in der Anwenderbroschüre für das Statistikpaket R beschrieben, da dieses bei Anwendern und Experten eine weit verbreitete, frei verfügbare Software ist.

Die Anwenderbroschüre enthält im ersten Kapitel Informationen zur Installation von R und MLPowSim sowie eine Beschreibung der zur Arbeit mit MLPowSim wesentlichen Parameter. Dies sind allgemeine Angaben, Informationen zu Hierarchieebenen und verfügbaren Designs, Informationen zur Modellierung der Kovariaten sowie die Angabe, ob für die Power zusätzlich Konfidenzintervalle berechnet werden sollen. Die allgemeinen Angaben umfassen Informationen zu Fehler 1. Art, Anzahl der Simulationen, Seed für Zufallszahlen, minimale und maximale Anzahl der Einheiten auf jeder Hierarchieebene, sowie Schrittweite. In Abhängigkeit der Skala der Zielgröße sind noch Schätzverfahren (bei metrischen Zielgrößen) bzw. Linkfunktion und Approximationsmethode zu spezifizieren (bei binären Zielgrößen). Das darauf folgende Kapitel enthält Details zur Powerberechnung mit MLPowSim. Der dann folgende Hauptteil der Anwenderbroschüre enthält je ein Kapitel zur Anwendung der Software für eine metrische und für eine binäre Zielgröße. Die Anwendung ist jeweils an einem konkreten Beispiel demonstriert. In jedem Kapitel ist zuerst der Beispieldatensatz beschrieben. Anschließend erfolgt eine Beschreibung der für das Beispiel notwendigen MLPowSim Eingabeparameter. Mithilfe von Screenshots wird dem Leser gezeigt, an welchen Stellen die Parameter eingegeben werden müssen. Das von MLPowSim erzeugte R-Programm heißt powersimu.r und ist für jedes der beiden Beispiele im Anhang der Anwenderbroschüre enthalten. Dieses Programm muss im Statistikpaket R geöffnet und ausgeführt werden. Im Ergebnis der Ausführung entsteht eine Textdatei powerout.txt, welche das Resultat der Simulationen enthält. Da die Monte-Carlo-Simulationen für diskrete Settings erfolgen (d.h. Kombinationen aus möglichen Anzahlen von Einheiten auf den verschiedenen Hierarchieebenen), besteht das Resultat aus Schätzwerten der statistischen Power für diese Settings. Neben dem Inhalt der Textdatei powerout.txt ist in der Anwenderbroschüre jeweils ein Ausschnitt aus einem der von powersimu.r erzeugten Datensätze beschrieben.

Formelsammlung

Als weiteres Ergebnis der Literatur- und Internetrecherchen wurde eine Reihe von Publikationen zum Thema „Fallzahl/Power bei komplexen hierarchischen Studiendesigns" gefunden und im Hinblick auf die Projektziele ausgewertet. Im Mittelpunkt stand die Suche nach Formeln zur Abschätzung von Power oder Fallzahl. Als Resultat wurden für einige spezielle komplexe hierarchische Studiendesigns entsprechende Formeln gefunden, die in einer Formelsammlung zusammengefasst wurden. Diese Studiendesigns sind cluster-randomisierte Studien im Parallelgruppendesign mit zwei oder drei Hierarchieebenen (metrische Zielgrößen) bzw. zwei Hierarchieebenen (binäre Zielgrößen). Jedem Design ist in der Formelsammlung ein eigener Abschnitt gewidmet. Jeder Abschnitt enthält zunächst einen theoretischen Teil, in welchem die Formeln erläutert werden. Der sich daran anschließende Teil enthält ein praktisches Beispiel, an welchem die Anwendung der Formeln demonstriert ist. Hierbei werden dieselben Beispiele wie in der Anwenderbroschüre verwendet. Die in Anwenderbroschüre und Formelsammlung enthaltenen Beispiele stammen aus realen Studien (Brandstetter et al. 2012; Dreyhaupt et al. 2012; Dreyhaupt et al. 2013b) und wurden für dieses Nachwuchsakademieprojekt noch erweitert und aufbereitet.

Im Vergleich zu MLPowSim hat die Formelsammlung den Vorteil einer vergleichsweise einfachen Anwendung. Der Nachteil ist,

dass Formeln nur für eine eingeschränkte Anzahl Studiendesigns verfügbar sind. Wie auch für die Anwendung von MLPowSim sind für die Anwendung der Formeln Vorinformationen aus Pilotstudien oder ähnlichen Projekten notwendig.

Informationen zu Power- und Fallzahlabschätzung für Situationen ohne verfügbare Vorinformationen

Der dritte Bericht, welcher im Rahmen dieses Nachwuchsakademieprojekts entstand, enthält Hinweise zur Power- und Fallzahlabschätzung im Kontext komplexer hierarchischer Datenstrukturen für den Fall, dass keine Vorinformationen verfügbar sind. Die Inhalte sind Arain et al. 2010; Eldridge/Kerry 2012; Lake et al. 2002; Lancaster et al. 2004; NETSCC 2014; Thabane et al. 2010 und van Schie/Moerbeek 2014 entnommen. Im Bericht wird zwischen internen und externen Pilotstudien sowie Feasibility-Studien unterschieden. Bei internen Pilotstudien muss die Power- und Fallzahlberechnung zusammen mit der Hauptstudie erfolgen, und damit für das primäre Zielkriterium der Hauptstudie. Bei externen Pilotstudien und Feasibility-Studien hängt die Fallzahl von den konkreten Studienzielen ab. Im Bericht sind daher nur Hinweise gegeben, was Ziele in externen Pilotstudien und Feasibility-Studien sein können. Spezielle Hinweise werden zu Konfidenzintervallen für Raten und für die Standardabweichung gegeben, da diese Parameter oft in solchen Studien geschätzt werden. Eine Powerberechnung für das Finden eines Unterschiedes in einer primären Zielgröße sollte in externen Pilotstudien und Feasibility-Studien nicht erfolgen (Eldridge/Kerry 2012).

Nutzung von Anwenderbroschüre und Formelsammlung bei metrischer Zielgröße

Beschreibung des Beispiels

Ein für die Versorgungsforschung wichtiger Bereich ist die Beurteilung der Wirksamkeit von Präventionsmaßnahmen, siehe z.B. Wright et al. 2014. Für die Planung einer solchen Studie sollen im Folgenden sowohl mit der Anwenderbroschüre als auch mit der Formelsammlung Power- bzw. Fallzahlberechnungen demonstriert werden. Die zu planende Studie wird zunächst charakterisiert.

In Anlehnung an die Baden-Württemberg-Studie (Dreyhaupt et al. 2012) soll eine landesweite zweiarmige longitudinale clusterrandomisierte Studie geplant werden (zwei Zeitpunkte: Baseline und Follow Up), in welcher die Wirksamkeit eines schulbasierten Präventionsprogramms untersucht werden soll. Randomisiert werden sollen Grundschulen in Baden-Württemberg in eine Interventionsgruppe bzw. eine Kontrollgruppe (1:1 Randomisierung). Als primäre Zielgröße soll bei den einzelnen Grundschulkindern die Änderung der Anzahl der gelaufenen Meter beim 6-Minuten-Lauf gemessen werden (Follow Up-Baseline). Da zwischen den Messwerten bei Kindern innerhalb einer Schule Abhängigkeiten bestehen können, ist bei der Studienplanung von einem hierarchischen Design auszugehen: Von einem Studiendesign mit zwei Hierarchieebenen (Schüler sind innerhalb der Grundschulen verschachtelt). Es wird weiterhin angenommen, dass in jeder Grundschule dieselbe Anzahl an Kindern in die Studie eingeschlossen werden kann. Bei diesem Design handelt es sich um ein balanciertes Studiendesign mit zwei Hierarchieebenen. Aus Gründen der Übersichtlichkeit soll in diesem Beispiel nur eine Kovariate betrachtet werden (Gruppenzugehörigkeit: Interventions- oder Kontrollgruppe) sowie die Einschränkung auf zwei Hierarchieebenen. Der Fehler 1. Art wird mit 0.05 (zweiseitig) angenommen. Eine Rekrutierung von insgesamt minimal 60 bis maximal 100 Schulen ist für die zu planende Studie realistisch. Für die Anzahl der Schüler innerhalb der Grundschulen wird angenommen, dass in jeder Schule 20, 30 oder 40 Schüler in die Studie aufgenommen werden können. Als Vorinformationen stehen Daten einer Pilotstudie zur Verfügung (URMEL-ICE Studie, Brandstetter et al. 2012). Bei der Studienplanung wird von einer zu erreichenden Power von 80 % bzw. 90 % ausgegangen.

Simulationen zur Powerabschätzung				
Setting	**Gesamtanzahl Schulen**	**Anzahl Schüler pro Schule**	**Gesamtanzahl Schüler**	**Power**
1	60	20	1200	76
2	60	30	1800	82
3	60	40	2400	86
4	80	20	1600	88
5	80	30	2400	92
6	80	40	3200	94
7	100	20	2000	94
8	100	30	3000	97
9	100	40	4000	98

Tab. 1: Simulationen zur Powerabschätzung: Charakteristik der 9 Settings und Ergebnis der Powersimulationen für das Beispiel.

Nutzung der Anwenderbroschüre

Für die Simulationen zur Powerabschätzung soll weiterhin angenommen werden, dass entweder 60, 80 oder 100 Schulen einbezogen werden können, woraus sich insgesamt 9 Settings (d.h. Kombination aus Schulanzahl und Anzahl Schüler) ergeben, für welche die Powerabschätzung über Simulationen erfolgen soll (siehe die ersten vier Spalten in Tabelle 1). Bei der Spezifizierung der Settings in MLPowSim muss die Gesamtanzahl an Schulen und die Anzahl Schüler pro Schule vorgegeben werden.

Zunächst wurde an die Daten der Pilotstudie das folgende lineare gemischte Regressionsmodell angepasst: $Y_{ij}=\beta_0+\beta_1 \cdot X_{1i}+b_{0i}+\varepsilon_{ij}$ (1)

Es bedeuten:

- Y_{ij}: Wert der primären Zielgröße beim j-ten Kind in der i-ten Schule (Änderung der Anzahl der gelaufenen Meter beim 6-Minuten-Lauf)
- X_{1i}: Gruppenzugehörigkeit der i-ten Schule ($X_{1i}=0$: Kontrollgruppe, $X_{1i}=1$: Interventionsgruppe)
- b_{0i}: zufälliger Intercept (Abweichung der i-ten Schule vom Gesamt-Intercept)
- ε_{ij}: Restfehler
- β_0 und β_1: Regressionskoeffizienten (Gesamt-Intercept und Effekt der Gruppenzugehörigkeit auf die Änderung der Anzahl der gelaufenen Meter beim 6-Minuten Lauf).

Die Regressionskoeffizienten β_0 und β_1 werden auch als feste Effekte bezeichnet, die Parameter b_{0i} und ε_{ij} als Varianzkomponenten (zufällige Effekte). Die Anpassung des oben beschriebenen linearen gemischten Regressionsmodells (1) an die Pilotdaten erfolgte mit dem Statistikpaket R (mit der Funktion lmer). Für die Modellparameter wurden folgende Schätzwerte erhalten: $\beta_0=21.109$, $\beta_1=29.550$, $Var(b_{0i})=994.6$, $Var(\varepsilon_{ij})=15242.8$.

Neben diesen Schätzwerten und den oben genannten allgemeinen Parametern sind für die Anwendung von MLPowSim für die Beispielstudie noch weitere Informationen notwendig, siehe Abschnitt 1.4.2 in der Anwenderbroschüre. Das im Ergebnis der Anwendung von MLPowSim entstandene R-Programm powersimu.r wurde in dem Statistikpaket R geöffnet und ausgeführt. Für jedes der 9 Settings wurde im Programmablauf eine Monte-Carlo-Simulation mit einer Vielzahl von Datensätzen ausgeführt. Im Ergebnis des Programmlaufs wurde für jedes Setting eine Power für den Gruppeneffekt erhalten (in der Textdatei powerout.txt gespeichert), siehe letzte Spalte in Tabelle 1.

Die Resultate zeigen, dass eine Power von mehr als 80 % mit 60 Schulen (d.h. 30 Schulen pro Gruppe) je 30 Schülern erreicht werden kann (insgesamt 1.800 Schüler). Eine Power von mehr als 90 % kann z.B. mit 80 Schulen (d.h. 40 Schulen pro Gruppe) je 30 Schülern oder 100 Schulen (d.h. 50 Schulen pro Gruppe) mit je 20 Schülern erreicht werden (insgesamt 2.400 Schüler oder 2.000 Schüler). Bei komplexen hierarchischen Studiendesigns ist es oft so, dass eine bestimmte Power (z.B. 80%) mit verschiedenen Settings erreicht wird. Hier muss der Anwender entscheiden, welches Setting in Bezug auf die zur Verfügung stehenden Ressourcen am ehesten machbar ist (hier: mehr Schüler pro Schule in weniger Schulen oder mehr Schulen mit weniger Schülern). Sollen noch weitere Kovariaten einbezogen (z.B. Alter, Geschlecht) oder eine Verfeinerung der Clusterstruktur vorgenommen werden (hier: zusätzliche Hierarchieebene durch Berücksichtigung von Klassen innerhalb Schulen), muss das lineare gemischte Regressionsmodell (1) entsprechend verändert und erneut an die Pilotdaten angepasst werden. Die Auswertung der Studie sollte später mit demselben linearen gemischten Regressionsmodell erfolgen, wie vorher in der Planung verwendet wurde (Dang et al. 2008).

Nutzung der Formelsammlung

Die Anwendung der Formelsammlung für die Planung der Beispielstudie ist im Vergleich zur Anwenderbroschüre einfacher, da alle Berechnungen z.B. mit einem Taschenrechner erfolgen können. Für cluster-randomisierte Studien im Parallelgruppendesign mit zwei Hierarchieebenen sind entsprechende Formeln in Abschnitt 1.1.1 der Formelsammlung verfügbar. Grundlage ist zunächst die für eine individuell randomisierte Studie berechnete Fallzahl (bei einer

Literatur

Adams, J.L./Wickstrom, S.L./Burgess M.J./Lee, P.P./Escarce, J.J. (2003): Sampling patients within physician practices and health plans: multistage cluster samples in health services research. In: Health Serv Res 2003 Dec;38(6 Pt 1):1625-40

Arain, M,/Campbell, M.J./Cooper, C.L./Lancaster, G.A. (2010): What is a pilot or feasibility study? A review of current practice and editorial policy. In: BMC Med Res Methodol 2010, 10: 67

Bell, B.A./Morgan, G.B./Schoeneberger, J.A./Loudermilk, B.L./Kromrey, J.D./Ferron,
J.M. (2010): Dancing the sample size limbo with mixed models: How low can you go?, SAS Global Forum 2010 Posters Paper 197-2010. In: http://support.sas.com/resources/papers/proceedings10/197-2010.pdf (zugegriffen am 27. Februar 2015)

Brandstetter, S./Klenk, J./Berg, S./Galm, C./Fritz, M./Peter, M./Prokopchuk, D./Steiner, R./Wartha, O./Steinacker, J.M./Wabitsch, M. (2012): Overweight prevention implemented by primary school teachers: a randomised controlled trial. In: Obes Facts 2012, 5(1): 1-11

Browne, W.J./Lahi, M.G./Parker. R.M.A. (2009): A Guide to Sample Size Calculations for Random Effect Models via Simulation and the MLPowSim Software Package.
School of Clinical Veterinary Sciences University of Bristol In: http://www.bristol.ac.uk/cmm/software/mlpowsim/ (zugegriffen am 27.Februar.2015)

Dang, Q./Mazumdar, S./Houck, P.R. (2008): Sample size and power calculations based on generalized linear mixed models with correlated binary outcomes. In: Comput Methods Programs Biomed 2008, Aug, 91(2): 122-127

Donner, A./Klar, N. (2010): Design and Analysis of Cluster Randomization trials in Health Research. Chichester: John Wiley & Sons, Ltd.

Dreyhaupt, J./Koch, B./Wirt, T./Schreiber, A./Brandstetter, S./Kesztyues, D./Wartha, O./Kobel, S./Kettner, S./Prokopchuk, D./Hundsdoerfer, V./Klepsch, M./Wiedom,
M./Sufeida, S./Fischbach, N./Muche, R./Seufert, T./Steinacker, J.M. (2012): Evaluation of a health promotion program in children: Study protocol and design of the cluster-randomized Baden-Wuerttemberg primary school study [DRKS-ID: DRKS00000494]. In: BMC Public Health 2012, 12: 157 [URL: http://www.biomedcentral.com/1471-2458/12/157]

Dreyhaupt, J./Sufeida, S./Steinacker, J.M./Muche, R. für die Arbeitsgruppe „Komm mit in das gesunde Boot – Grundschule“ (2013a): Power- und Fallzahlabschätzungen für hierarchische und longitudinale Studien. In: Muche, R./Minkenberg, R. (2013): KSFE 2013 – Proceedings der 17. Konferenz der SAS®-Anwender in Forschung und Entwicklung (KSFE). Aachen: Shaker Verlag: 111-120

Dreyhaupt, J./Sufeida, S./Steinacker, J.M./Muche, R. (2013b): Fallzahl- und Powerabschätzung für komplexe Studiendesigns in der Epidemiologie. In: Abstractband der 8. Jahrestagung der DGEpi und 1. Internationales LIFE Symposium, Leipzig 2013: 12-13

Eldridge, S./Kerry, S. (2012): A Practical Guide to Cluster Randomised Trials in Health Services Research. New York: Wiley

Fitzmaurice, G.M./Laird, N.M./Ware, J.H. (2011): Applied Longitudinal Analysis, Second Edition. Hoboken, NJ, USA John Wiley & Sons, Inc.

Hey, M./Maschewsky-Schneider, U. (2006): Kursbuch Versorgungsforschung. Berlin: MWV

Heo, M./ Leon, A.C. (2008): Statistical Power and Sample Size Requirements for Three Level Hierarchical Cluster Randomized Trials. In: Biometrics 2008, 64: 1256- 1262.

Karbach, U./ Stamer, M./ Holmberg, C./ Güthlin, C./ Patzelt, C./ Meyer, T. (2012): Qualitative Studien in der Versorgungsforschung - Diskussionspapier, Teil 2. Stand qualitativer Versorgungsforschung in Deutschland - ein exemplarischer Überblick. In: Gesundh.wes 2012, 74: 516-525

Kontopantelis, E./Springate. D.A./Parisi, R./Reeves, D. (2014): Simulation-based power calculations for mixed effects modelling: ipdpower in Stata. In: Journal of Statistical Software 2014 (Online), eScholarID:233423 In: https://www.escholar.manchester.ac.uk/uk-ac-man-scw:233423 (zugegriffen am 27. Februar 2015)

Lake, S./Kammann, E./Klar, N./Betensky, R. (2002): Sample size re-estimation in cluster randomization trials. In: Stat Med 2002, May 30, 21(10): 1337-1350 Lancaster, G.A./Dodd, S./Williamson, P.R. (2004): Design and analysis of pilot studies: recommendations for good practice. In: J Eval Clin Pract 2004, May, 10(2): 307-312

MLwiN Homepage: http://www.bristol.ac.uk/cmm/software/mlwin/ (zugegriffen am 10. April 2015)

Meinck, S./Vandenplas, C. (2012): Sample size requirements in HLM: An empirical study. In: IERI Monograph Series Issues and Methodologies in Large-Scale Assessments. IER Institute, Special Issue 1, Educational Testing Service and International Association for the Evaluation of Educational Achievement.

Meyer, T./ Karbach, U./ Holmberg, C./ Güthlin, C./ Patzelt, C./ Stamer, M. (2012): Qualitative Studien in der Versorgungsforschung - Diskussionspapier, Teil 1: Gegenstandsbestimmung. In: Gesundh.wes 2012, 74: 510-515

NETSCC NIHR Evaluation, Trials and Studies Coordinating Centre. Glossary. In: http://www.nets.nihr.ac.uk/glossary (Abruf am 27. Februar 2015)

Pfaff, H./Glaeske, G./Neugebauer, E.A.M./Schrappe, M. (2009): Memorandum III: Methoden für die Versorgungsforschung, Teil 1. In: Gesundh.wes 2009, 71: 505-510 Pfaff, H./Neugebauer, E.A.M./Glaeske, G./Schrappe, M. (Hrsg.) (2011): Lehrbuch Versorgungsforschung: Systematik-Methodik-Anwendung. Stuttgart: Schattauer Verlag

R-Homepage: http://www.r-project.org/ (zugegriffen am 10. April 2015)

Reich, N.G./Myers, J.A./Obeng, D./Milstone, A.M./Perl, T.M. (2012): Empirical power and sample size calculations for cluster-randomized and cluster-randomized crossover studies. In: PLoS One 2012, 7(4):e35564. doi: 10.1371/journal.pone.0035564

Thabane, L./Ma, J./Chu, R./Cheng, J./Ismaila, A./Rios, L.P./Robson, R./Thabane, M./Giangregorio, L./Goldsmith, C.H. (2010): A tutorial on pilot studies: the what, why and how. In: BMC Med Res Methodol 2010, Jan 6,10: 1. doi: 10.1186/1471-2288-10- 1

van Schie, S./Moerbeek, M. (2014): Re-estimating sample size in cluster randomised trials with active recruitment within clusters. In: Stat Med 2014, Aug 30, 33(19): 3253- 3268

Wright, D.R./Taveras, E.M./Gillman, M.W./Horan, C.M./Hohman, K.H./Gortmaker, S.L./Prosser, L.A. (2014): The cost of a primary care-based childhood obesity prevention intervention. In: BMC Health Serv Res 2014, Jan 29, 14: 44

Zhong Y., Cook R.J. (2015): Sample size and robust marginal methods for cluster- randomized trials with cen-sored event times. In: Stat Med 2015, Mar 15;34(6): 901- 923. doi: 10.1002/sim.6395. Epub 2014 Dec 17

vorgegebenen Power), die dann mit dem Variationsinflationsfaktor (Designeffekt DE) korrigiert wird (siehe z.B. Eldridge/Kerry 2012). In den Designeffekt gehen Informationen über Clustergröße und Abhängigkeiten der Messwerte innerhalb der Cluster ein (Intraclass-Korrelationskoeffizient (ICC)). Wie auch bei der Anwenderbroschüre sind hier Vorinformationen notwendig. Hier werden Schätzwerte der Modellparameter aus der Anpassung des linearen gemischten Regressionsmodells (1) an die Daten einer Pilotstudie verwendet: β_0=21.109, β_1=29.550, Var(b_{0i})=994.6, Var(ε_{ij})=15242.8.

Aus diesen Daten ergibt sich eine Standardabweichung von $\sqrt{994.6 + 15242.8}$ = 127.43 und ein ICC von 994.6/(15242.8 + 994.6) = 0.061. Die Berechnung der Fallzahl pro Gruppe für eine individuell randomisierte Studie kann beispielsweise nach Formel (7.1) in Eldridge/Kerry 2012 vorgenommen werden: $N=2\cdot(z_{1-\alpha/2} + z_{1-\beta})^2\cdot\sigma^2/(\mu_1-\mu_2)^2$, mit $z_{1-\alpha/2}$ bzw. $z_{1-\beta}$ ist das 1-α/2 Quantil bzw. 1-β Quantil der Standardnormalverteilung für einen zweiseitigen Fehler 1. Art von α und eine Power von 1-β.

Mit μ_1 und μ_2 sind die arithmetischen Mittelwerte in den beiden Gruppen bezeichnet, σ ist die gemeinsame Standardabweichung der Zielgröße. Mit den obigen Angaben können diese Parameter wie folgt belegt werden: μ_1 = 21.109 und μ_2 = 21.109 + 29.550 = 50.659. Für α = 0.05 (zweiseitig) und 1-β= 0.80 bzw. 0.90 erhält man $z_{0.975}$ = 1.96 und $z_{0.80}$ = 0.84 bzw. $z_{0.90}$ = 1.28.

Die Fallzahl pro Gruppe in einer individuell randomisierten Studie ergibt sich

- für eine Power von 80 % aus $2\cdot(1.96+0.84)^2\cdot127.43/(21.109-50.659)^2$ = 291.59, d.h. pro Gruppe 292 Kinder, insgesamt 584 Kinder
- für eine Power von 90 % aus $2\cdot(1.96+1.28)^2\cdot127.43/(21.109-50.659)^2$ = 390.43, d.h. pro Gruppe 391 Kinder, insgesamt 782 Kinder.

Diese Anzahlen werden nun um den Designeffekt korrigiert: DE = 1+ICC·(m-1), mit der festen Clustergröße m (hier im Beispiel die Anzahl der Kinder pro Schule) und dem aus den Daten berechneten ICC. Wie in der obigen Nutzung der Anwenderbroschüre wird für die Anzahl der Schüler innerhalb der Grundschulen angenommen, dass in jeder Schule entweder 20, 30 oder 40 Schüler in die Studie aufgenommen werden können. Damit ergeben sich drei Werte für den Designeffekt: DE_{20} = 1+0.061·19 = 2.159, DE_{30} = 1+0.061·29 = 2.769 und DE_{40} = 1+0.061·39 = 3.379.

Die mit dem Designeffekt multiplizierten Fallzahlen zeigt Tabelle 2. Für die Berechnung wurde zunächst die mit dem Designeffekt korrigierte Gesamtanzahl an Schülern berechnet, dann davon ausgehend die Anzahl der Schulen, wobei davon ausgegangen wurde, dass diese durch zwei teilbar sein soll (1:1 Randomisierung). Die Spalte ‚Gesamtanzahl Schüler' enthält das Produkt Anzahl Schulen x Anzahl Schüler pro Schule.

Gesamtanzahlen nach Anwendung der Formelsammlung				
Anzahl Schüler pro Schule	Power 80%		Power 90%	
	Gesamtanzahl Schüler	Gesamtanzahl Schulen	Gesamtanzahl Schüler	Gesamtanzahl Schulen
20	1.280	64	1.680	84
30	1.620	54	2.160	72
40	2.000	50	2.640	66

Tab. 2: Gesamtanzahlen Schüler und Schulen nach Anwendung der Formelsammlung.

Im Folgenden werden die Ergebnisse am Beispiel der Settings mit m=30 Schülern pro Schule verglichen. Im Vergleich mit der Monte-Carlo-Simulation (Anwenderbroschüre, siehe oben) entspricht am ehesten das Setting 2 in Tabelle 1 (insgesamt 60 Schulen mit je 30 Schülern; 1800 Schüler in der Gesamtstudie) mit einer Power von 82 % dem Ergebnis von 54 Schulen mit je 30 Schülern in Tabelle 2. Eine erneute Simulation mit MLPowSim für 54 Schulen mit 30 Schülern liefert eine Power von 79%, was sehr nah an den hier zugrunde liegenden 80 % liegt.

Mit Setting 5 in Tabelle 1 (insgesamt 80 Schulen mit je 30 Schülern; 2400 Schüler in der Gesamtstudie) wird bei den Monte-Carlo-Simulationen eine Power von 92 % erreicht. Diesem Setting entspricht in Tabelle 2 am ehesten das Ergebnis von 72 Schulen mit je 30 Schülern bei einer Power von 90%. Eine erneute Simulation mit MLPowSim für 72 Schulen mit 30 Schülern liefert eine Power von 89%, was den angestrebten 90 % sehr nah kommt. Somit sind die Resultate der Nutzung der Anwenderbroschüre und der Formelsammlung für das Beispiel vergleichbar. Die Unterschiede sind u.a. durch Rundungsfehler zu begründen. Leicht veränderte Fallzahlen für die individuell randomisierte Studie können auch auftreten, wenn die Berechnung auf anderem Weg, beispielsweise mittels Spezialsoftware (z.B. nQuery, PROC power in SAS) realisiert wird. Anwenderbroschüre, Formelsammlung und der Bericht mit Informationen zu Power- und Fallzahlabschätzung für Situationen ohne verfügbare Vorinformationen können als gemeinsames Dokument auf der Internetseite http://vts.uni-ulm.de/query/longview.meta.asp?document_id=9509 heruntergeladen werden.

Diskussion/Fazit

Im Ergebnis dieses Projekts wurden drei Berichte erstellt, die als Instrumente zur Planung und Bewertung konfirmatorischer Studien mit metrischen und binären Zielgrößen bei komplexen hierarchischen Studiendesigns in der Versorgungsforschung und anderen Gebieten dienen können. Insbesondere mit der in der Anwenderbroschüre beschriebenen Nutzung von MLPowSim kann eine Vielzahl komplexer hierarchischer Studiendesigns mit bis zu drei Hierarchieebenen behandelt werden. Dies dürfte für viele praktische Anwendungen genügen. Die Anwenderbroschüre basiert auf der Verwendung der Spezialsoftware MLPowSim zur Erstellung von Quellcode für ein Simulationsprogramm in der Statistiksoftware R. Neben Kenntnissen in MLPowSim (zur korrekten Spezifizierung der Parameter) sind für die Nutzung der Anwenderbroschüre auch Statistikkenntnisse und Grundkenntnisse in der Statistiksoftware R hilfreich. Für die Simulationen werden mit MLPowSim zunächst diskrete Settings definiert, für welche dann in R eine Monte-Carlo-Simulation erfolgen kann. Die Power wird dann für die diskreten Settings erhalten. Ein Nachteil ist, dass MLPowSim für Studien mit mehr als drei Hierarchieebenen oder longitudinale Studiendesigns nicht verwendet werden kann. In diesen Situationen muss ein eigenes Simulationsprogramm erstellt werden.

Die Formelsammlung kann für einige spezielle komplexe hierarchische Studiendesigns verwendet werden. Ihre Anwendung ist vergleichsweise einfach, da alle Berechnungen z.B. mit einem Taschenrechner erfolgen können. Allerdings ist ihre Verwendung auf relativ wenige Studiendesigns eingeschränkt. Im Gegensatz zur Anwenderbroschüre wird hier nicht von vorher zu definierenden

diskreten Settings ausgegangen, sondern auf Basis eines vorgegebenen klinischen Mindesteffektes für eine festgelegte Power (z.B. 80 % oder 90%) zunächst die Fallzahl für eine individuell randomisierte Studie berechnet. Sowohl zwischen den Ergebnissen der Monte-Carlo-Simulationen und der Formelanwendung als auch bei der Anwendung verschiedener Formeln in der Formelsammlung können teilweise leichte Unterschiede auftreten (je nach verwendeten Ansatz und Annahmen). Die im Rahmen der Studienplanung angewendeten Methoden sollten daher auch für die statistische Auswertung verwendet werden.

Mit der Verfügbarkeit von Formeln für weitere Situationen kann die Formelsammlung künftig noch erweitert werden. Sowohl Anwenderbroschüre als auch Formelsammlung können zusätzlich noch um Zähldaten (z.B. Donner/Klar 2010: 64-65; Eldridge/Kerry 2012: 145-146) oder Ereigniszeiten (Zhong/Cook 2015) ergänzt werden.

Eine wesentliche Limitation von Anwenderbroschüre und Formelsammlung besteht in der Notwendigkeit des Vorliegens geeigneter Vorinformationen aus Pilotstudien, ähnlichen Projekten oder der Literatur. Falls diese Informationen nicht verfügbar sind, muss zunächst eine Feasibility- oder Pilotstudie durchgeführt werden, wofür der dritte Bericht Hinweise enthält.

Die Versorgungsforschung wird von den drei vorgestellten Instrumenten insbesondere profitieren, da in diesem Gebiet nicht nur cluster-randomisierte Studien eine zentrale Rolle spielen, sondern auch die Berücksichtigung von Confoundern immer wichtiger wird. <<

Tool for power and sample size estimation in complex hierarchical study designs in health services research

Background: Studies are an important source of knowledge in health service research. Here, designs with hierarchical data structures and additional consideration of covariates (complex hierarchical data structures) often occur. Examples are longitudinal studies in which the time course of certain characteristics is measured (e.g. body weight) and additional covariates are considered (e.g. sex, age at baseline) or cluster randomized studies with consideration of further cluster characteristics (e.g. percentage of migrants, baseline values for the outcome). In case of confirmatory studies in such situations, a-priori power or sample size calculation is needed but available methods are limited.
Methods: Procedures/formulas for sample size and power calculation in situations with complex hierarchical data structures were established via extensive searches.
Results: A freeware for creating source code for power calculation via simulations was found. This software can be used both for continuous and binary endpoints in many typical study situations in health services research. A report (user brochure) was developed for a user-friendly application of this software. Formulas for sample size calculation in some special situations were summarized in a second report (formulary). A third report contains information for planning and carrying out feasibility or pilot studies in situations with complex hierarchical data structures. The user brochure and the formulary are exemplary used for planning of a study with continuous outcome which is described in the article.
Discussion/Conclusions: The three reports provide a supporting tool for power and sample size calculation in studies with complex hierarchical study structures.

Keywords

hierarchical study designs with covariates, power and sample size calculation, simulation, cluster randomized trials

Dr. Jens Dreyhaupt
ist wissenschaftlicher Mitarbeiter am Institut für Epidemiologie und Medizinische Biometrie der Universität Ulm. Der Beitrag entstand im Rahmen des Nachwuchsprogramms des Netzwerks „Versorgungsforschung Baden-Württemberg“.

Kontakt: jens.dreyhaupt@uni-ulm.de

