Daten aus zweiter Hand

Datenreanalyse zur Überprüfung explorativer Hypothesen in der psychologischen Forschung

Martin J. Binser, Matthias Spörrle und Friedrich Försterling

Ludwig Maximilians Universität München

Abstract

Zielsetzung
Eignungsprüfung der Methode:
Knowledge Discovery in Databases (KDD)
für die psychologische Forschung

Definition KDD
Identifizierung von Mustern in großen Datensätzen

Methode
Vergleich der Befunde von KDD mit klinischer Studie bzgl. Depression nach fötalem Verlust

Ergebnis
Relatives Risiko: KDD = 2.1; klinische Studie = 2.9

Diskussion
- Ökonomie bei Passung des Primärdatensatzes
- zur Testung explorativer Hypothesen geeignet

kumulierter Zeitaufwand

Einleitung

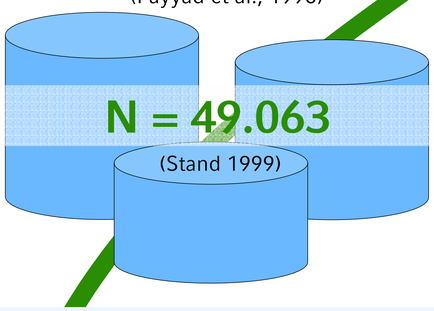
Ausgangslage
Zunehmende Verfügbarkeit „psychologischer" Primärdaten (z.B. Primärdatenbanken, Datenverarbeitung einer Krankenkasse)

Fragestellung
KDD als valide, reliable, objektive und ökonomische Methode in der psychologischen Forschung?

Hypothese
KKD kann ein erhöhtes relatives Risiko für Depression nach fötalem Verlust nachweisen.

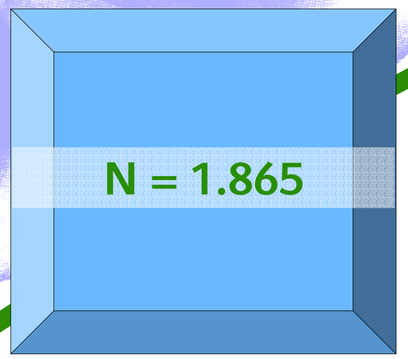
Methode

(Fayyad et al., 1996)



Daten(bank)auswahl

Medical Expenditure Panel Survey
(**MEPS**; Cohen, 1997)
- frei verfügbare medizinische Langzeitstudie der amerikanischen Gesundheitsbehörde
- repräsentativer Durchschnitt der USA
- reliable Messung durch CAPI und Daten der Versicherungsträger
- Angaben über Fehlgeburten (Befragung & ICD-9)
- Angaben über die Behandlung psychischer Erkrankungen (ICD-9)



Datenaufbereitung

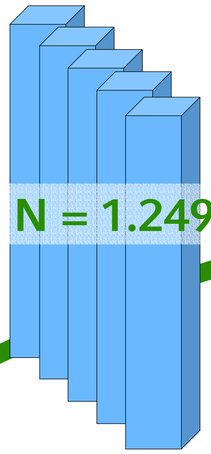
Einschlusskriterium (N = 1.865)
nur Frauen mit Schwangerschaftsereignis (Fehlgeburt, Totgeburt, Geburt)

Problem
- Informationen in über 100 Datenfiles verteilt (verschiedene Jahre, medizinische Information nach Ort der Behandlung, etc.)

MEPS: überlappendes Paneldesign in fünf Runden über 2 Jahre hinweg

MEPS	Kalenderjahr			
Panel	1996	1997	1998	1999
	Runde 1A 2A 3A 4A 5A			
1 (A)	22.598 Personen	20.866 Pers.		
		Runde 1B 2B 3B 4B 5B		
2 (B)		13.679 Pers.	12.933 Pers.	
			Runde 1C 2C 3C 4C 5C	
3 (C)			11.439 Pers.	10.439 Pers.

Lösung
- Bündelung der Info in einem Arbeitsfile
- Variablenberechnung in div. Originalfiles



Daten-vorverarbeitung

Probleme
- starke Heterogenität der Daten, z.B. Geburts-ereignisse: taggenau (74%) bis 5 Monate
- Informationen zum Schwangerschaftsstatus z.T. inkonsistent (AHRQ, 2003)
- unklare zeitliche Assoziation von Depression und Schwangerschaftsereignis

Lösung (=> N = 1.249)
- Reliabilitätsindex für Schwangerschaftsereig. Ausschluss: Vertrauensintervalle > 3 Monate
- Validitätsindex für Schwangerschaftsereig. Ausschluss z.B. berichtete Geburt ohne Kind
- Reduktion des Auswertungsniveaus: Dichotomisierung, da keine zeitliche Validierung der Depression möglich

=> **Resultierende Variablen:**
Depression (dichotom)
- mindestens ein Depressionsereignis (ICD-9) in 2 J. (gesamter Beobachtungszeitraum)

Fötaler Verlust (dichotom)
- mindestens eine Fehl- oder Totgeburt in 2 Jahren

Literatur
- Agency for Healthcare Research and Quality. (2003). *MEPS HC-046 1996-2000 Pregnancy Files.* Retrieved 25-03, 2002, from www.meps.ahrq.gov available on demand
- Cohen, S. (1997). Sample design of the 1996 Medical Expenditure Panel Survey Household. *MEPS Methodology Report, 2,* Pub. No. 97-0027.
- Fayyad, U.-M., Piatetsky-Shapiro, G., Smyth, P., & Uthurusamy, R. (Eds.). (1996). *Advances in knowledge discovery and data mining.* Menlo Park.
- Neugebauer, R., Kline, J., O'Connor, P., Shrout, P., Johnson, J., Skodol, A., et al. (1992). Depressive symptoms in women in the six months after miscarriage. *American Journal of Obstetrics and Gynecology, 166*(1 Pt 1), 104-109.

PDF dieses Posters unter www.binser.de

Relatives Risiko (RR)

2.1 KDD
2.9 Klinisch

Muster (Ergebnis)

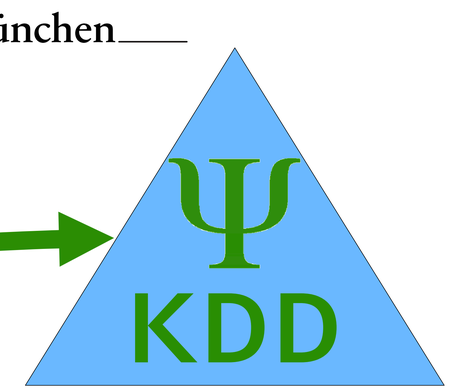
Absolute Häufigkeiten
- 17,7% aller Schwangerschaften fötale Verluste (incl. 1,1% Totgeburten)
- 4,4% depressive Erkrankung (6,9% insgesamt psychische Erkrankungen)

Auswahl klinische Studie Neugebauer et al. (1992)
- größte klinische Studie zum Thema
- Schwangere als Kontrollgruppe
- Maß CES-D als Schätzung für die Häufigkeit depressiver Erkrankungen

KKD (N = 1249) vs. klinische Studie (N = 515)

	Relatives Risiko	Abs. Häufig. föt. Verlust	Abs. Häufig. Geburt	χ^2
KDD	**2.1**	7.7%	3.7%	6.79**
Klinische Studie	**2.9**	24.1%	8.3%	23.82***

$^{**}p < .01$, zweiseitig
$^{***}p < .001$, zweiseitig



Wissen (Diskussion)

Validität
+ KDD guter Schätzer für Verhältnisse (RR)
– Sicherung der Validität (Depression) führt zu Verlust des Datenniveaus
– keine Kausalitätsaussage möglich (Depression, Folge oder Auslöser?)

Reliabilität
+ nach Ausschlüssen hohe Reliabilität
– aber unter Verringerung der Stichprobe

Ökonomie
+ Zeitaufwand Eigenerhebung > KDD
– Datenauswahl, -aufbereitung und Datenvorverarbeitung verursacht hohen Zeitaufwand!

Objektivität
+ möglich bei Reduktion auf geringes Datenniveau (hier Dichotomisierung)
– aufwendig z.B. bei Bestimmung präziser zeitlicher Assoziationen von Ereignissen

Fazit
- wenn Primärdatensatz mit hoher Passung zur Hypothese vorhanden
- geeignete Methode zur Exploration psychologischer Hypothesen

Primärdatenbanken im Internet

www.nsd.uib.no/cessda
Council Of European Social Science Data Archives

www.psychdata.zpid.de (im Aufbau)
Zentrum für Psychologische Information und Dokumentation

www.gesis.org/za
Zentralarchiv für Empirische Sozialforschung der Universität Köln