

# Einsatz Statistischer Verfahren bei Benchmarkingprozessen in der Versorgungsforschung

ein methodischer Beitrag zur Analyse von Registerdaten

vorgelegt von  
Diplom-Statistiker Harald Siedentop  
aus Berlin

von der Fakultät VII - Wirtschaft und Management der  
Technischen Universität Berlin zur Erlangung des akademischen Grades  
Doktor der Gesundheitswissenschaften  
- Dr. P.H. -

genehmigte Dissertation

Promotionsausschuss:

Vorsitzender: Prof. Dr. Reinhard Busse  
Berichter: Prof. Dr. M. Harvey Brenner  
Prof. Dr. Karl Wegscheider

Tag der wissenschaftlichen Aussprache: 20. Juni 2008

Berlin 2008

D 83

## **Danksagungen:**

Mein Dank gilt Herrn Prof. Dr. M Harvey Brenner und PD Dr. Susanne Dahms (†) für die Betreuung dieser Arbeit sowie Prof. Dr. Karl Wegscheider für die Übernahme der Begutachtung nach dem Ableben von Frau Dahms.

Desweiteren danke ich dem Berliner Herzinfarktregister e.V. für die Erlaubnis zur Datennutzung und Herrn Dipl-Math. Gerd Kallischnigg für seine Hilfe bei der Themenstellung.

---

# Inhaltsverzeichnis

<b>1</b>	<b>Einleitung</b>	<b>1</b>
1.1	Problemstellung . . . . .	1
1.2	Zielstellung dieser Arbeit . . . . .	4
1.3	Anwendungsbeispiel . . . . .	5
1.4	Bemerkungen zur Erhebungsmethode . . . . .	6
1.5	Ranking-Systeme in der Praxis . . . . .	8
1.5.1	Ziele . . . . .	8
1.5.2	Beispiele für bestehende Ranking-Systeme im Ausland . . . . .	9
1.5.2.1	Das USNWR Ranking System . . . . .	9
1.5.2.2	Das AHRQ Hospital Rating . . . . .	14
1.5.2.3	Krankenhaus-Ranking beim THCIC in Texas . . . . .	16
1.5.2.4	Vergleiche zwischen Bypass Chirurgen (NY) . . . . .	17
1.5.3	Situation in Deutschland . . . . .	18
1.6	Ranking-Systeme und Öffentliche Gesundheit . . . . .	19
1.6.1	Interviews mit Gesundheitsexperten . . . . .	19
1.6.2	Beitrag aus der Beratungswirtschaft . . . . .	20
1.7	Ranking-Systeme aus statistischer Sicht . . . . .	21
1.8	Diskussion der Literatur und Schlussfolgerung . . . . .	24
<b>2</b>	<b>Statistische Methodik</b>	<b>25</b>
2.1	Einführung . . . . .	25
2.2	Klassische Lineare Modelle . . . . .	29
2.2.1	Varianzanalyse . . . . .	30
2.2.1.1	Einfache Varianzanalyse, Einfachklassifikation . . . . .	30
2.2.1.2	Blockexperimente . . . . .	39
2.2.1.3	Zweifache Varianzanalyse . . . . .	42

2.2.2	Regressionsanalyse . . . . .	44
2.2.3	Kovarianzanalyse . . . . .	45
2.2.4	Generalisierte Lineare Modelle – Logistische Regressionsanalyse	46
2.2.5	Poisson-Modelle . . . . .	55
2.3	Gemischte Lineare Modelle . . . . .	56
2.3.1	Einführung . . . . .	56
2.3.2	Varianzkomponenten-Modelle . . . . .	62
2.3.2.1	Beispiel: Die einfache hierarchische Klassifikation . .	63
2.3.3	Varianzkomponenten-Modelle mit Kovariaten . . . . .	68
2.3.4	Modelle mit zufälligen Koeffizienten . . . . .	68
2.3.5	Varianzexzess bei gemischten linearen Modellen . . . . .	68
2.3.5.1	Motivation, Auswirkung zufälliger Effekte . . . . .	68
2.3.5.2	Minimierung des Prognosefehlers (BLUP) . . . . .	74
2.3.5.3	Empirische Eigenschaften von Modell-Schätzern . . .	80
2.3.5.4	Betrachtungen zur Testgüte . . . . .	85
2.3.5.5	Klinik-spezifische Konfidenzintervalle und Tests . . .	88
2.4	Generalisierte Gemischte Lineare Modelle . . . . .	92
2.4.1	Einführung . . . . .	92
2.4.2	Streuungsursachen bei generalisierten gemischten Modellen . .	95
2.4.2.1	Motivation, Auswirkung zufälliger Effekte . . . . .	95
2.4.2.2	Modellschätzer und Tests im Random-Logit-Modell .	100
2.4.2.3	Ein Signifikanztest zur Globalhypothese . . . . .	104
2.4.2.4	Zentrums-spezifische Konfidenzintervalle und Tests .	106
2.5	Alternative Analyseverfahren . . . . .	112
2.6	CART Verfahren . . . . .	113
2.6.1	Motivation . . . . .	113
2.6.2	Modellbildung . . . . .	118
2.6.3	Modellschätzung . . . . .	120
2.6.3.1	Bestimmung von Split-Regeln . . . . .	121
2.6.3.2	Bestimmung von Stopp-Regeln . . . . .	127
2.6.4	Diskussion und Fazit . . . . .	128
<b>3</b>	<b>Analysestrategie</b>	<b>130</b>
3.1	Wahl der statistischen Modellklasse . . . . .	130
3.2	Variablenselektion zur Risikoadjustierung . . . . .	131

3.2.1	Auswahl über Datenexploration, robuste Verfahren . . . . .	131
3.2.2	Auswahl nach Signifikanz, Selektionsverfahren . . . . .	132
3.2.3	Auswahl nach Einflussgrad . . . . .	134
3.2.4	Auswahl nach medizinisch-fachlichen Gesichtspunkten . . . . .	134
3.3	Test-Prozedur . . . . .	135
3.4	Umgang mit fehlenden und unplausiblen Werten . . . . .	136
3.5	Darstellung der Ergebnisse . . . . .	137
3.6	Anmerkungen zur Analyse-Software . . . . .	138
<b>4</b>	<b>Anwendungsbeispiel: Das BHiR</b>	<b>140</b>
4.1	Medizinische Grundlagen . . . . .	140
4.1.1	Anatomie und Physiologie . . . . .	141
4.1.2	Epidemiologie und Burden of Disease . . . . .	142
4.2	Vorbetrachtungen . . . . .	145
4.2.1	Beschreibung des Patientenkollektivs . . . . .	145
4.2.2	Charakteristika der Datenbank . . . . .	147
4.2.2.1	Erfasste Parameter . . . . .	147
4.2.2.2	Plausibilitätsprüfung . . . . .	148
4.2.2.3	Fehlende Werte . . . . .	149
4.3	Nicht adjustierte Analysen . . . . .	150
4.3.1	Separate Betrachtung der Kliniken . . . . .	152
4.3.2	Nicht adjustierte Modellierung nach Klinikum im GLM . . . . .	153
4.3.3	Übergang zum Generalisierten Gemischten Linearen Modell . . . . .	156
4.4	Kovariaten zur Risikoadjustierung . . . . .	159
4.4.1	Exkurs: Auswahl nach Signifikanz (Demographische Variablen)	160
4.4.1.1	Vorbemerkungen zur Modellbildung . . . . .	161
4.4.1.2	Logistische Regressionsanalyse . . . . .	162
4.4.1.3	CART-Analyse für demographische Variablen . . . . .	172
4.4.1.4	Univariate Betrachtungen . . . . .	175
4.4.1.5	Bivariate Betrachtungen (Patientenalter) . . . . .	180
4.4.1.6	Schlussfolgerung . . . . .	186
4.4.2	Auswahl nach fachlichen Gesichtspunkten . . . . .	187
4.5	Modellaufbau zur Risikoadjustierung . . . . .	191
4.5.1	Parametrischer Ansatz (logistische Regressionsanalyse) . . . . .	191
4.5.2	CART-Analyse zur Exploration geeigneter Risikofaktoren . . . . .	193

4.6	Ergebnisse zum Einrichtungsvergleich . . . . .	197
4.6.1	Hauptmodell, primäre Analyse im adjustierten GLMM . . . . .	197
4.6.2	Ergebnisse anderer Modellierungen / Robustheitsprüfung . . . . .	199
4.6.2.1	GLMM mit vollständiger Erfassung . . . . .	200
4.6.2.2	GLMM, andere Modelle . . . . .	201
4.6.2.3	Ein klassisches Generalisiertes Lineares Modell . . . . .	203
4.6.3	Fazit . . . . .	205
<b>5</b>	<b>Zusammenfassung und Diskussion</b>	<b>207</b>
<b>A</b>	<b>Literatur</b>	<b>213</b>
<b>B</b>	<b>Nebenrechnungen</b>	<b>225</b>
B.1	Rechnung zur Variation der binomialverteilten Zentrumshäufigkeiten	225
B.2	Rechnung zum Impurity-Maß . . . . .	227
<b>C</b>	<b>Verwendete Analyseprogramme (Auszüge)</b>	<b>233</b>
C.1	SAS Programmcodes zur Simulation / LMM . . . . .	233
C.1.1	Programm zum Makro-Aufruf (balancierter Fall) . . . . .	233
C.1.2	Simulationsmakro . . . . .	233
C.2	SAS Programmcodes zur Simulation / GLMM . . . . .	236
C.2.1	Programm zum Makro-Aufruf (unbalancierter Fall, BHiR) . . . . .	236
C.2.2	Simulationsmakro . . . . .	236
C.3	SAS Programmcode zur Datenaufbereitung und Datenersetzungen . . . . .	239
C.4	SAS Programmcode für primäres Analysemodell . . . . .	240
C.5	R Codes zum Aufruf von CART-Analysen . . . . .	241
C.5.1	Analyse demographischer Variablen (Exkurs) . . . . .	241
C.5.2	Parameterselektion zur Hauptanalyse . . . . .	241

---

# Tabellenverzeichnis

2.1	Einfache ANOVA Tafel . . . . .	33
2.2	Tafel des einfachen Blockexperiments . . . . .	41
2.3	Tafel der einfachen hierarchischen Klassifikation, balancierter Fall . .	66
2.4	Tafel der einfachen hierarchischen Klassifikation, unbalancierter Fall .	67
2.5	Simulationsergebnisse EBLUP-Schätzer, balancierter Fall . . . . .	82
2.6	Simulationsergebnisse EBLUP-Schätzer, unbalancierter Fall . . . . .	85
2.7	Signifikanzniveau des Wald-Z-Tests für $\sigma_a^2$ . . . . .	86
2.8	Signifikanzniveau des Wald-Z-Tests für $\sigma_a^2$ , viele Zentren . . . . .	86
2.9	Simulationsergebnisse zu Klinik-spezifischen Konfidenzintervallen . .	90
2.10	Simulation zur Variabilität der Häufigkeiten, Random-Logit-Modell .	102
2.11	Simulation für Modellschätzer im Random-Logit-Modell . . . . .	103
2.12	Simulation zu Klinik-spezifischen Tests im Random-Logit-Modell, $H_0$	108
2.13	Simulation zu Klinik-spezifischen Tests im Random-Logit-Modell, $H_1$	109
4.1	Krankheitskosten 2004 je Einwohner nach Krankheitsklassen und Alter	144
4.2	Patienten-Stichprobe des BHiR . . . . .	146
4.3	Beobachtete Krankenhaussterblichkeit nach Klinikum . . . . .	151
4.4	Krankenhaussterblichkeit nach Klinikum und Jahr . . . . .	152
4.5	Krankenhaussterblichkeit nach Klinikum, nicht adjustiertes GLM . .	154
4.6	Krankenhaussterblichkeit nach Klinikum, nicht adjustiertes GLMM .	157
4.7	Logistische Regression 1 – demographische Variablen . . . . .	163
4.8	Vollständigkeit der Erhebung und Mortalität . . . . .	164
4.9	Logistische Regression 2 – demographische Variablen . . . . .	166
4.10	Logistische Regression 3 – demographische Variablen . . . . .	167
4.11	Letalität nach Altersgruppen und Geschlecht, vollständige Fälle . . .	176
4.12	Letalität nach Altersgruppen und Geschlecht, alle verfügbaren Fälle .	176

4.13 Letalität nach BMI-Klassen . . . . .	177
4.14 Letalität nach Nationalität . . . . .	179
4.15 Letalität nach Wohnort . . . . .	179
4.16 Alter (in Jahren) nach Geschlecht . . . . .	182
4.17 Body Mass Index nach Altersklassen . . . . .	182
4.18 Alter (in Jahren) nach Nationalität . . . . .	183
4.19 Alter (in Jahren) nach Wohnort . . . . .	184
4.20 Alter (in Jahren) nach Familienstand . . . . .	185
4.21 Alter (in Jahren) nach Berufsgruppe . . . . .	186
4.22 Risikofaktoren (1) nach Klinikum . . . . .	189
4.23 Risikofaktoren (2) nach Klinikum . . . . .	190
4.24 Modellbildung zur Risikoadjustierung (feste Effekte) . . . . .	192
4.25 Ergebnisse für adjustiertes Auswertungsmodell nach Schritt 4, GLMM	198
4.26 Ergebnisse für adjustiertes Auswertungsmodell nach Schritt 1, GLMM	200
4.27 Ergebnisse für adjustiertes Auswertungsmodell, nach CART, GLMM	203
4.28 Ergebnisse für adjustiertes Auswertungsmodell nach Schritt 4, GLM	204

---

# Abbildungsverzeichnis

2.1	Simulationsergebnisse zum Varianz-Exzess unter $H_1$ . . . . .	71
2.2	Simulationsergebnisse zum Varianz-Exzess unter $H_0$ . . . . .	72
2.3	Simulationsergebnisse zum Varianz-Exzess, unbalancierter Fall . . . . .	73
2.4	Auswirkung der Schrumpfung auf Effekt-Schätzer . . . . .	77
2.5	Auswirkung negativer Schätzung der Zentrums-Variabilität . . . . .	78
2.6	Mittlere Prediktionsfehler bei normalverteilten Mittelwerten . . . . .	84
2.7	Güte des Wald-Z-Tests für $\sigma_a^2$ , balancierter Fall . . . . .	87
2.8	Mischverteilung von binomialverteilten Zufallsgrößen . . . . .	96
2.9	Simulationsergebnisse zu Z-Statistiken im Random-Logit-Modell, $H_0$ .	105
2.10	Empirische Verteilung der EBLUP-Schätzer, gegeben $H_0$ . . . . .	110
2.11	Empirische Verteilung der EBLUP-Schätzer, gegeben $H_1$ . . . . .	111
2.12	Darstellung eines Klassifikations- bzw. Regressionsbaums . . . . .	115
2.13	Impurity-Maße bei binärer Zielgröße . . . . .	125
2.14	Impurity-Maße bei drei Ausprägungen der Zielgröße . . . . .	126
4.1	Herzkammer nach einem Herzinfarkt (Kernspintomographie) . . . . .	142
4.2	Krankenhaussterblichkeit nach Klinikum, separate Betrachtung . . . . .	151
4.3	Krankenhaussterblichkeit nach Klinikum, GLM . . . . .	155
4.4	Krankenhaussterblichkeit nach Klinikum, GLMM . . . . .	158
4.5	Krankenhaus-Letalität nach Altersgruppe . . . . .	171
4.6	CART-Analyse für demographische Variablen . . . . .	173
4.7	Letalität nach Altersklassen, Daten und Analyse-Ergebnisse . . . . .	174
4.8	Zusammenhang des Alters mit demographischen Kovariaten . . . . .	181
4.9	CART-Analyse für mögliche Risikofaktoren, volles Modell . . . . .	194
4.10	CART-Analyse für mögliche Risikofaktoren, reduziertes Modell . . . . .	196
4.11	Letalität nach Klinikum, GLMM adjustiert . . . . .	199

4.12	Effekt-Schätzer für Kliniken, GLMM adjustiert (Modelle 1 bis 6) . . .	201
4.13	Adjustiertes Modell 4, Effekt-Schätzer im GLM und GLMM . . . . .	205

---

# Abkürzungsverzeichnis

AHRQ	Agency for <u>H</u> ealth <u>C</u> are <u>R</u> esearch and <u>Q</u> uality
AHT	<u>A</u> rterielle <u>H</u> ypert <u>o</u> nie)
AMA	<u>A</u> merican <u>M</u> edical <u>A</u> ssociation
ANOVA	Varianzanalyse ( <u>A</u> nalysis of <u>V</u> ariance)
AOK	<u>A</u> llgemeine <u>O</u> rts- <u>K</u> rankenkassen
BHiR	<u>B</u> erliner <u>H</u> erzinfarkt <u>R</u> egister
BLUE	<u>B</u> est <u>L</u> inear <u>U</u> nbiased <u>E</u> stimation
BLUP	<u>B</u> est <u>L</u> inear <u>U</u> nbiased <u>P</u> rediction
BMI	<u>B</u> ody <u>M</u> ass <u>I</u> ndex [ $kg/m^2$ ]
BQS	<u>B</u> undesgeschäftsstelle <u>Q</u> ualitäts <u>S</u> icherung
CART	<u>C</u> lassification <u>A</u> nd <u>R</u> egression <u>T</u> rees
CMS	<u>C</u> enters for <u>M</u> edicare and Medicaid <u>S</u> ervices
df	Freiheitsgrade (degrees of <u>f</u> reedom)
DHZB	<u>D</u> eutsches <u>H</u> erzzentrum <u>B</u> erlin
DM	<u>D</u> iabetes <u>M</u> ellitus
EBLUE	<u>E</u> mpirical <u>B</u> est <u>L</u> inear <u>U</u> nbiased <u>E</u> stimation
EBLUP	<u>E</u> mpirical <u>B</u> est <u>L</u> inear <u>U</u> nbiased <u>P</u> rediction
EBS	<u>E</u> mpirical <u>B</u> ayes <u>S</u> chätzer
EF	<u>E</u> jektions <u>f</u> raktion
ETG	<u>E</u> mpirical <u>T</u> riple <u>G</u> oal
GLM	<u>G</u> eneralisiertes <u>L</u> ineares <u>M</u> odell
GLMM	<u>G</u> eneralized <u>L</u> inear <u>M</u> ixed <u>M</u> odel
HCE	<u>H</u> yper <u>ch</u> olesterinämie
IHQ	<u>I</u> ndex of <u>H</u> ospital <u>Q</u> uality
KHK	<u>K</u> oronare <u>H</u> erzer <u>k</u> rankung
KI	<u>K</u> onfidenz <u>i</u> ntervall

LOR	Log Odds Ratio
LP	Linearer Prediktor (im GLM bzw. GLMM)
LSB	Linksschenkelblock
MHI	Manifeste Herzinsuffizienz
MINQUE	Minimum Variance Quadratic Unbiased Estimation
ML	Maximum Likelihood
MMPL	Maximum Marginal Pseudo Likelihood
MS	Mean Square (mittlere Quadratsumme)
MSPL	Maximum Subject-Specific Pseudo Likelihood
NI	Niereninsuffizienz
NORC	National Organization for Research, Chicago
OR	Odds Ratio
PA	Population Average
PSI	Patient Safety Indicators
QSR	Qualitätssicherung der stationären Versorgung mit Routinedaten
RD	Risiko-Differenz (risk difference)
REML	Restricted Maximum Likelihood
RMPL	Residual Marginal Pseudo Likelihood
RPL	Restricted Pseudo Likelihood
RR	Relatives Risiko (relative risk)
RSPL	Residual Subject-Specific Pseudo Likelihood
SD	Standard Deviation (Standardabweichung)
SEM	Standard Srror of the Mean (Standardfehler des Mittelwerts)
SS	Sum of Squares (Quadratsumme)
TG	Triple Goal
THCIC	Texas Health Care Information Council
USNWR	United States News & World Report
Z.n.I	Zustand nach früherem Infarkt

---

# 1 Einleitung

## 1.1 Problemstellung

Im Bereich der Gesundheitsversorgung ist häufig eine Situation anzutreffen, in der für eine bestimmte medizinische Fragestellung eine Vielzahl von Therapieoptionen verfügbar ist. Ein Patient ist daran interessiert, nach Möglichkeit die für ihn beste Therapie zu erhalten. In vielen Fällen ist die Therapiestrategie an die Wahl des Behandlungszentrums, also einer Klinik oder einer Praxis, gebunden. Das bedeutet, dass der Patient wissen möchte, wo er die beste Behandlung erhält.

Die Erfolgsaussichten einer Behandlung können von unterschiedlichen weiteren Faktoren abhängen, wie beispielsweise

- prognostische Faktoren, die der Patient aufweist (z.B. soziodemographische oder anamnestische Variablen);
- technische Ausstattung der jeweiligen medizinischen Praxis bzw. der medizinischen Klinik;
- Entfernung vom Wohnort des Patienten zur medizinischen Einrichtung;
- Erfahrung des medizinischen Personals bezüglich der spezifischen Indikation (z.B. Anzahl von durchgeführten Operationen);
- Ausbildungsstand des medizinischen Personals vor Ort (etwa Weiterbildungs-Politik).

Oft ist es nicht möglich, die beste Klinik oder Praxis zu erreichen, selbst wenn die genannten Faktoren vollständig bekannt und aufbereitet sind. Dafür können folgende Gründe eine Rolle spielen:

- Notfallindikationen, bei denen die Transportzeit zur nächstliegenden Einrichtung (Notaufnahme) eine entscheidende Rolle spielt oder keine Zeit vorhanden ist, eine Wahl des Behandlungszentrums zu treffen;
- das „beste“ Behandlungszentrum liegt zu weit entfernt;
- die Diagnose ist unklar;
- Informationen über Qualitätsmerkmale der zur Wahl stehenden Einrichtungen sind unbekannt.

Es kann somit zu sehr unterschiedlichen Erfolgsaussichten der jeweils gewählten Therapie für den einzelnen Patienten kommen, weil Patienten unter Umständen nicht der für ihre individuelle Situation optimalen Therapie zugeführt werden können. In Notfallindikationen, wie beispielsweise beim akuten Myokardinfarkt, kann dies ernsthafte Folgen haben.

Es ist zu untersuchen, ob nicht gerade in Ballungsräumen, wo häufig für eine einzelne Indikation eine größere Anzahl von medizinischen Zentren in geringer Entfernung und damit aufgrund des heterogenen Ausstattungs- und Ausbildungsstands eine Fülle von Therapiemöglichkeiten zur Verfügung steht, eine frühzeitige und individuell abgestimmte Entscheidung hinsichtlich der zu wählenden Klinik sinnvoll ist.

Gerade unter Notfallbedingungen sind Leitlinien für die Behandlung, die beispielsweise anhand von bestimmten anamnestischen bzw. patientenbezogenen Kenngrößen festgelegt werden können, wünschenswert. Diese Leitlinien könnten bereits bei der Wahl des Klinikums – in Abhängigkeit der geographischen Lage und der technischen Ausstattung – eingesetzt werden.

Häufig ist jedoch die beste Therapiemethode für die individuelle Kombination aus Indikation und bestehenden Begleiterkrankungen unbekannt. In solchen Situationen ist ohne empirische Daten keine Verbesserung der Versorgungssituation denkbar.

Für viele Indikationen existieren umfangreiche Datenbestände in Form von REGISTERN. Unter einem Register versteht man allgemein eine systematische Sammlung von Informationen über eine Gruppe von Objekten. Im Bereich des Gesundheitswesens sind Register somit Fallsammlungen zu Patienten mit einer bestimmten Eigenschaft (z.B. mit einer Indikation wie Krebserkrankung oder Myokardinfarkt).

Diese Sammlungen sind in der Regel auf ein bestimmtes Gebiet und auf bestimmte Einrichtungen beschränkt. Man unterscheidet bei Registern außerdem zwischen *epidemiologischen* und *klinischen* Registern. In epidemiologischen Registern untersucht man Faktoren, die zur Erkrankung oder Nicht-Erkrankung der Subjekte beitragen; bei klinischen Registern steht der Behandlungserfolg bei Vorliegen einer bestimmten Erkrankungen im Vordergrund.

Für die Indikation des akuten Myokardinfarkts existiert beispielsweise ein solches klinisches Register. Im Rahmen des „Berliner Herzinfarktregister e.V.“ (BHiR) [63] wurde ein umfassender Datenbestand von Infarkt-Patienten gesammelt, die innerhalb eines bestimmten Zeitraums in eines der teilnehmenden Zentren eingeliefert wurden. Aufgezeichnet wurde unter anderem

- die behandelnde Notfallklinik;
- demographische Daten;
- Angaben zu präexistierenden Erkrankungen;
- Riskofaktoren;
- Angaben zur gewählten Notfallbehandlung;
- durchgeführte Wiederbelebungsmaßnahmen;
- die gewählte anschließende Therapie;
- Komplikationen;
- der Therapieerfolg, definiert durch das Überleben des Patienten (ja oder nein) während des Klinikaufenthalts.

## 1.2 Zielstellung dieser Arbeit

Das Ziel dieser Arbeit ist die Entwicklung einer Auswertungsstrategie für vorhandene und zukünftige Registerdaten. Von primärem Interesse ist hierbei eine Benchmark-Bildung für Behandlungszentren, die im Rahmen dieser Arbeit wie folgt verstanden wird:

BENCHMARKING bezeichnet die Identifikation von Qualitätsmerkmalen und die Klassifizierung von Behandlungszentren hinsichtlich des Therapieerfolgs bei einer oder mehrerer medizinischen (Sub)Indikation(en). Bei der Klassifizierung von Zentren soll der Therapieerfolg unter Berücksichtigung von prognostischen Faktoren der behandelten Patienten gemessen werden.

Zur Benchmarkbildung finden sich in der Praxis häufig – basierend auf Registerdaten – sogenannte „Rankings“ oder „Ranking-Systeme“, mittels derer die Behandlungsqualität von Einrichtungen einer Untersuchungseinheit (etwa einer Region oder eines Typs) gemessen und in eine Rangfolge gestellt wird. Im Rahmen dieser Arbeit sollen die Begriffe „Ranking“ und „Einrichtungsvergleich“ synonym verwendet werden.

Eine generelle Eigenschaft von Registerdaten ist, dass es sich in der Regel um nicht kontrollierte und nicht randomisierte Versuche handelt. Das bedeutet, dass die Subjekte (hier die Patienten) den Einrichtungen nicht zufallsgesteuert oder geschichtet („stratifiziert“) und damit annähernd gleichmäßig verteilt zugeordnet werden, sondern nach vielen anderen Kriterien, wie etwa der geographischen Lage, der Ausstattung der Einrichtung oder auch dem Schweregrad einer Erkrankung. Es handelt sich somit eher um Klumpenstichproben („clustered data“). Dies bedeutet statistisch, dass die Beobachtungen innerhalb der Zentren voneinander abhängig sein können, da Einrichtungen möglicherweise Patientengruppen mit einrichtungsspezifischen Eigenschaften behandeln oder behandelt haben. Um einen objektiven Vergleich zwischen den Einrichtungen zu gewährleisten, ist eine geeignete Adjustierung der Ergebnisse hinsichtlich prognostischer Risikofaktoren erforderlich.

Im Rahmen dieser Arbeit werden statistische Verfahren gegenübergestellt und diskutiert, mit denen Datenbestände hinsichtlich der oben beschriebene Problemstellung analysiert werden können. Bei Vergleichen zwischen Behandlungszentren

soll untersucht werden, inwieweit sich die Wahl der Analysemethode auf die Beurteilung des Therapieerfolgs auswirkt. Gesundheitsökonomische Kenngrößen (wie etwa Lebensqualität, Kosten-Nutzwert oder Risiko-Nutzwert) können ggf. aus den Ergebnissen abgeleitet werden. Etablierte statistische Analysemethoden sind unter anderem

- klassische lineare Modelle (wie logistische Regressionsanalyse oder Varianzanalyse);
- mehrstufige Verfahren;
- Varianzkomponenten-Modelle;
- hierarchische Verfahren;
- CART-Analysen (Classification and Regression Trees).

Die Eigenschaften der zur Verfügung stehenden Methoden werden anhand von Beispielen und Simulationen diskutiert. Aus den Ergebnissen werden dann Empfehlungen für zukünftige Datenerhebungen bzw. Verfahren zur Beurteilung von Benchmarking-Prozessen abgeleitet und beispielhaft auf den Datenbestand des Berliner Herzinfarktregisters angewendet werden. Der Schwerpunkt dieser Arbeit liegt in der Methodenbetrachtung, woran sich die Auswertung des Beispiel-Datenbestandes und dessen Ergebnis-Interpretation untermauernd anschließt.

## 1.3 Anwendungsbeispiel

Eine mögliche Anwendungs-Situation liegt in der Betrachtung des Behandlungserfolgs, der gemessen werden kann als Mortalität von Herzinfarkt-Patienten, bis 30 Tage nach Auftreten des Infarkts. Diese dichotome oder binäre Zielgröße (nur zwei Ausprägungen sind möglich: Erfolg bzw. Misserfolg) kann von der Behandlungsstrategie, den anamnestischen Faktoren des Patienten, von sonstigen Charakteristika des Zentrums (Entfernung, technische Ausstattung, Patientenaufkommen etc.) oder von der Zeitspanne, die zwischen Infarkt und Beginn der Akutbehandlung vergangen ist, abhängen.

Ziel der Analyse ist es, die Frage zu beantworten, ob zwischen Behandlungszentren *generell* oder in der vorliegenden Datenbasis selbst unterschiedliche Chancen bestehen, einen Therapieerfolg hinsichtlich der betrachteten Zielgröße zu erzielen. Es soll somit untersucht werden, ob aufgrund der in der Stichprobe beobachteten Unterschiede auf tatsächliche Unterschiede zwischen den Zentren in der Grundgesamtheit geschlossen werden kann. Diese und weitere Fragen (z.B. welche Faktoren besitzen Einfluss auf den Behandlungserfolg) schlüssig beantworten zu können, soll die im Methodenteil zu identifizierende geeignete Analysestrategie auf den vorliegenden Beispieldatensatz angewendet werden.

## 1.4 Bemerkungen zur Erhebungsmethode

Günstige Voraussetzungen für den Nachweis möglicher Unterschiede zwischen den Ausprägungen prognostischer Faktoren, wie etwa Art der Behandlung oder Behandlungszentrum, sind durch die Wahl von prospektiv geplanten, randomisierten (und stratifizierten) Studien gegeben. Über den Ansatz der zufälligen und balancierten Zuordnung innerhalb von Stratifizierungsvariablen lässt sich am ehesten gewährleisten, dass sich die Subjekte und deren Kovariaten gleichmäßig über die Stufen verteilen, so dass die Fallzahlverhältnisse zwischen den Zellen annähernd konstant sind und ein unmittelbarer Vergleich zwischen den Stufen (auch bei zufälliger Ungleichverteilung von Kovariaten) zulässig ist. Dieser Ansatz wird üblicherweise bei klinischen Studien der Phasen II und III gewählt.

### Beispiel

Als illustrierendes Beispiel soll eine zwei-faktorielle Studie mit jeweils zwei Stufen pro Einflussfaktor benutzt werden.

Zum Vergleich von  $p$  Kliniken hinsichtlich des Behandlungserfolgs bei KHK-Patienten soll eine prospektive randomisierte Studie mit einer Gesamtzahl von  $N$  Patienten durchgeführt werden. Um Gleichverteilung zu erreichen, wird eine 1:1-Randomisierung, stratifiziert nach Geschlecht, welches als wichtigste Kovariate eingeschätzt wird, gewählt. Es ergeben sich entsprechend  $2p$  Zellen:

Behandlungszentrum	Geschlecht		Summe
	männlich	weiblich	
1	$n_{1m}$	$n_{1w}$	$n_{1.}$
2	$n_{2m}$	$n_{2w}$	$n_{2.}$
...	...	...	...
$i$ ( $i < p$ )	$n_{im}$	$n_{iw}$	$n_{i.}$
...	...	...	...
$p$	$n_{pm}$	$n_{pw}$	$n_{p.}$
Summe	$n_{.m}$	$n_{.w}$	$n_{..}$

Durch die nach Geschlecht stratifizierte Randomisierung wird sichergestellt, dass nicht nur die Bedingung

$$n_{i.} = n_{j.} \quad (i \neq j),$$

sondern auch die Forderungen

$$n_{im} = n_{jm} \quad (i \neq j) \quad \text{und} \quad n_{iw} = n_{jw} \quad (i \neq j),$$

also eine Balanciertheit zwischen den Zentren und jeweils innerhalb beider Geschlechtsgruppen, möglichst exakt erfüllt wird.

Ein weiteres Qualitätsmerkmal von kontrollierten Studien ist die so genannte „Verblindung“ der Behandlung. Soll etwa der Vorteil einer neuen Prüfsubstanz gegenüber einem Standardmedikament nachgewiesen werden, so muss die verabreichte Substanz sowohl dem behandelnden Arzt als auch dem Patienten unbekannt sein. Auf diese Weise werden subjektive Einflüsse bei der Beurteilung des Behandlungserfolgs weitestgehend ausgeschlossen.

In der Situation von klinischen Registerdaten sind diese Voraussetzungen in aller Regel nicht erfüllt. Die Patienten werden sich nicht durch einen zentralen Mechanismus gesteuert auf die Untersuchungseinheiten (die teilnehmenden Kliniken) verteilen, sondern bedingt durch viele andere Faktoren. Dies ist insbesondere bei der zugrunde liegenden Indikation des Anwendungsbeispiels dieser Arbeit, dem Berliner Herzinfarktregister, der Fall. Da der Behandlungserfolg, also das Überleben des Patienten, maßgeblich von der Zeitspanne bestimmt wird, die zwischen dem Auftreten des Notfalls (akuter Myokardinfarkt) und dem Erreichen der Klinik vergeht, muss

das Klinikum aufgrund der Nähe zum Wohnort bzw. Standort oder der schnellst möglichen Erreichbarkeit ausgewählt werden.

Bei der Betrachtung von klinischen Registerdaten geht es somit um nicht intervenierende Begleitung von laufenden Prozessen und deren Qualitätskontrolle und -verbesserung und nicht um prospektive Planung. Die hiermit einhergehenden methodischen Randbedingungen müssen durch die Wahl geeigneter statistischer Verfahren und Modelle berücksichtigt werden. Um Unterschiede zwischen den Patientenpopulationen in den teilnehmenden Zentren auszugleichen, bieten sich zum Beispiel Risikoadjustierungen an, bei der Faktoren, die vom Patienten getragen werden und den Behandlungserfolg beeinflussen können, in die statistische Modellierung einbezogen werden.

## 1.5 Ranking-Systeme in der Praxis

Viele Strategien bezüglich der Analyse von Registerdaten werden in der Literatur beschrieben. Einige werden im Abschnitt 1.5.2 beispielhaft beschrieben. Ihre Anfänge gehen auf Ansätze der britischen Krankenschwester Florence Nightingale [40] zurück, die 1863 Qualitätsdaten mit dem Ziel, die Hygienesituation in Feldlazaretten zu verbessern, sammelte.

### 1.5.1 Ziele

Die Etablierung von Ranking-Systemen im Gesundheitswesen dient heute der Messung und Darstellung der Behandlungs- bzw. Versorgungsqualität von Einrichtungen. Dabei werden verschiedene Ziele verfolgt:

- Qualitätssicherung für die Krankenhäuser;
- Vergleichbarkeit der Einrichtungen soll hergestellt werden;
- Transparenz: Qualitätsvergleiche zwischen Kliniken sollen der Öffentlichkeit zugänglich gemacht werden;
- Rankings können zur Vertragsgestaltung zwischen Leistungserbringern und -erstatern genutzt werden.

All diese Ziele sollen letztlich der Verbesserung der Gesundheitsversorgung dienen.

Als die zentrale Frage in dieser Arbeit wird erörtert, wie diese Rankings am besten erstellt werden können, so dass die Vergleiche zwischen den Einrichtungen valide, korrekt und reproduzierbar sind.

Die nachfolgenden Darstellungen zeigen einen Ausschnitt aus den gängigsten Publikationen zu Ranking-Systemen im Gesundheitswesen.

## 1.5.2 Beispiele für bestehende Ranking-Systeme im Ausland

Ein großer Teil der Methodenvorschläge und -diskussionen stammt aus den USA, da dort im Vergleich zu Europa ein wesentlich stärker ausgeprägter Wettbewerb zwischen den Leistungserbringern und damit eine stärkere Nachfrage nach Vergleichen besteht. Aber auch in europäischen Ländern wie Großbritannien, Frankreich oder Polen existieren verschiedene Ranking-Systeme.

### 1.5.2.1 Das USNWR Ranking System

Von der Zeitschrift *United States News & World Report* wird jährlich ein Krankenhaus-Ranking – bezogen auf 17 unterschiedliche Indikationsgruppen, wie beispielsweise bösartige Neubildungen, Herz-Kreislaufkrankungen oder urologische Erkrankungen – herausgegeben. Dieses Ranking beruht auf einer Analysemethode, die unter Federführung der *National Organization for Research der Universität Chicago (NORC)* entwickelt wurde [77]. Im Rahmen der Berichterstattung dieser Rankings werden unter dem Titel „America’s Best Hospitals“ für jede der Indikationsgruppen Zentren mit den besten Ranking-Ergebnissen ermittelt und benannt.

In die Auswertung werden dabei grundsätzlich alle etwa 6.000 Krankenhäuser in den USA einbezogen, jedoch kommen nur diejenigen Häuser tatsächlich zur Analyse, die bestimmte Einschlusskriterien erfüllen. Es verbleiben danach etwa 2.000 Zentren, die für wenigstens eine von 12 der insgesamt 17 Indikationsgruppen (Krebserkrankungen, Erkrankungen des Verdauungstrakts, HNO-Erkrankungen, Geriatrie, Gynäkologie, Herz-Kreislaufkrankungen, Hormonstörungen, Nierenerkrankungen, Neurologie, Orthopädie, Erkrankungen der Atmungsorgane, Urologie) auswertbar

sind. Beispielsweise ist für die Indikation des akuten Myokardinfarkts (welche auch dem Beispieldatensatz dieser Arbeit zugrunde liegt) im Jahr 2007 ein Bericht zum Klinikvergleich auf Basis der bis 2003 erfassten Daten erschienen [58].

### Zusammenfassung der Methodik:

Jedes im Sinne der Einschlusskriterien auswertbare Zentrum erhält einen Score-Wert, genannt „Index of Hospital Quality (IHQ)“, der aus drei Indikatoren der Behandlungsqualität zusammengesetzt ist:

- **Strukturelle Eigenschaften** des Zentrums, für die eine Reihe von Indizes definiert werden, von denen hier nur einige genannt seien:
  - Technologie-Index (für jede Indikationsgruppe separat): Für jede vor Ort vorhandene Technologie, die für eine bestimmte Indikationsgruppe relevant ist (zum Beispiel ein Computer-Tomographie-Scanner für Krebserkrankungen oder ein Herzkatheter-Labor für Herz-Kreislaufkrankungen), erhält das Zentrum einen Punkt. Zentren, die die jeweilige Technik über einen lokalen Dienstleister anbieten, erhalten hierfür einen halben Punkt.
  - Ein Volumen-Score, der pro Indikationsgruppe die Anzahl der erstattungsrelevanten Entlassungen reflektiert.
  - Das Zahlen-Verhältnis zwischen Krankenschwestern („full-time equivalents“) und den stationär sowie ambulant behandelten Patienten, das den Behandlungsaufwand pro Patient beschreibt.
- Der Entscheidungsprozess innerhalb eines Zentrums hinsichtlich einzusetzender Diagnostik oder Behandlungsmethoden, Überweisungen in andere Zentren oder Stationen sowie die Aufenthaltsdauer der Patienten werden mit dem Begriff „**prozessuale Eigenschaften**“ überschrieben. Da dieser Indikator sehr schwer zu messen bzw. zu quantifizieren ist, werden hierfür subjektive Maße, so genannte „Nominierungen“, die etwa durch klinische Experten erteilt werden, herangezogen. Wenn ein solcher Experte ein Klinikum als eines der besten einstuft, bekundet er damit seine Zustimmung zum dortigen Entscheidungsprozess.  
Um diese Nominierungen zu erheben, werden jährlich Umfragen durchgeführt,

deren Ergebnisse sich jeweils auf die letzten drei Jahre beziehen. Aus den insgesamt ca. 800.000 in der *American Medical Association* (AMA) eingetragenen Ärzten werden hierfür etwa 230.000 zertifizierte Ärzte ausgewählt, die bestimmte Auswahlkriterien erfüllen. Unter diesen wird eine nach Region und Fachgebiet innerhalb der Region geschichtete (stratifizierte) Zufallsstichprobe mit einem Umfang von ca. 2.500 Ärzten gezogen. Die für die Stichprobe ausgewählten Ärzte erhalten einen kurzen Fragebogen zugesandt, gefolgt von einer Erinnerungspostkarte, etwa eine Woche später. Weitere drei Wochen danach wird ein weiterer Erinnerungsbrief versandt. Die so erzielte Ausschöpfungsquote der Umfrage liegt bei etwa 50%, ein recht hoher Wert bei einer schriftlichen Befragung. In Deutschland liegen die Ausschöpfungsquoten bei dieser Art der Befragung üblicherweise im Bereich von 20% oder darunter.

Jeder Arzt, der den Fragebogen ausgefüllt hat, bekommt schließlich einen Gewichtungsfaktor – entsprechend seiner Auswahlwahrscheinlichkeit, die von der individuellen Schichtgröße abhängt – zugewiesen.

- Der dritte Indikator des Ranking-Scores misst die Behandlungsergebnisse (das „Outcome“). Unabhängig von der Indikationsgruppe wird in diesem Modell die **Mortalität** als (einzige) Zielgröße betrachtet, die nicht als Überlebenszeit („time-to-event“ Variable), sondern als Rate gemessen wird. Hierbei wird eine positive „Korrelation“ zwischen geringeren Mortalitätsraten und guter Behandlungsqualität postuliert.

Entscheidend zur Bildung des Mortalitäts-Score ist der Vergleich zwischen risikoadjustierten erwarteten und den tatsächlich beobachteten Sterberaten. Da die Fallzahlen und die Schweregrade der behandelten Patienten über die große Anzahl von Zentren stark variieren können, wird eine Risiko- und Schweregrad-Adjustierung der Mortalitätsraten durchgeführt, die – abhängig von der Indikationsgruppe – unter anderem folgende Variablen einschließt:

- Diagnose des Patienten bei Eintritt in die Klinik;
- durchgeführte Behandlungen;
- Alter;
- andere Komorbiditäten;
- andere Komplikationen;
- Wechselwirkungen zwischen verschiedenen Diagnosen.

Bei allen möglichen verschiedenen Diagnose- und Morbiditätskombinationen werden, basierend auf der gesamten Stichprobe, diejenigen Faktoren identifiziert, die einen signifikanten Beitrag zur Variabilität der Sterbewahrscheinlichkeit lieferten. Krankenhäuser, die einen höheren Anteil von Patienten mit höheren Risikostufen aufweisen, besitzen somit eine höhere erwartete Mortalitätsrate als Krankenhäuser, die eher Patienten mit niedrigem Risiko behandelten.

Basierend auf den Ausprägungen der Risikofaktoren wird für jeden Patienten eine skalare, ordinal skalierte Risikostufe, die einen Wert zwischen 1 (niedrig) und 4 (sehr hoch) annehmen kann, bestimmt. Jeder dieser vier Stufen wird eine feste Sterbewahrscheinlichkeit zugeordnet. Unter Berücksichtigung dieser Stufen kann für jedes teilnehmende Krankenhaus ( $h$ ) für eine Indikationsgruppe ( $i$ ) eine erwartete Mortalitätsrate  $E(d_{hi})$  (mit  $E()$ : Erwartungswert) bestimmt werden, die dem tatsächlich beobachteten Ergebnis  $d_{hi}$  gegenübergestellt wird und als Verhältnis  $R_{hi} = d_{hi}/E(d_{hi})$  angegeben wird. Werte unterhalb von 1 deuten auf eine im Vergleich zum Durchschnitt geringere Mortalität hin.

Diese Verhältnisse müssen schließlich noch in den **Mortalitäts-Score**, der positivere Werte für bessere Behandlungsqualität ausweisen soll, übersetzt werden. Dies geschieht durch Subtraktion des Wertes  $R_{hi}$  von der Zahl 1. Somit erhalten Zentren mit unterdurchschnittlicher Mortalitätsrate positive Score-Werte (maximal: 1), und „schlechtere“ Zentren negative Werte, die nach unten nicht beschränkt sind). Um kurzfristige Einflüsse zu eliminieren, werden jeweils die gleitende Durchschnitte aus den Score-Werten der letzten drei Jahre ausgewiesen.

Der Qualitätsindex wird schließlich wie folgt gebildet:

$$IHQ_i = \sum_{j=1}^J S_j F_j + P_i \sum_{k=1}^K + M \sum_{l=1}^L F_l,$$

wobei

$IHQ_i$  den Index for Hospital Quality für die jeweilige Indikationsgruppe  $i$ ,

$S_j$  die Struktur-Indikatoren,

$F_j$  die Gewichte der jeweiligen Struktur-Indikatoren,

$P_i$  den „Nominierungs-Score“ für Indikationsgruppe  $i$ , und

$M$  den standardisierten Mortalitäts-Score

bezeichnen.

Zum Zwecke der einheitlichen Darstellung werden diese Scores letztlich innerhalb der Indikationsgruppe  $i$  standardisiert und normiert, so dass das jeweilige Hospital mit dem höchsten  $IHQ_i$ -Score den Wert 100 erhält.

## Diskussion

Aufgrund des Wettbewerbssystems zwischen den Krankenhauseinrichtungen einerseits und der transparenten Berichterstattung der Zentren an zentrale Datenbanken und der daraus resultierenden umfangreichen Datenlage andererseits stehen dem US-amerikanischen Verbraucher Instrumente, wie das oben beschriebene USWNR Rating-System, zur Verfügung. Unabhängig von der Frage nach der Validität der Ergebnisse kann dieses System auf Deutschland nicht übertragen werden, da derart umfassende Datenbanken nicht existieren. Vielmehr bestehen hierzulande eher singuläre Registerdatenbanken, wie etwa das „Berliner Herzinfarktregister“ oder zahlreiche Krebsregister, bei denen primär Patientendaten bzw. klinische Ergebnisse erfasst werden.

Weiterhin stellt sich die Frage, ob tatsächlich weichere Faktoren, wie beispielsweise der Betreuungsaufwand (gemessen durch die Anzahl von Pflegepersonen je Patient) oder die technische Ausstattung – wie hier geschehen – als Zielgröße aufgefasst werden sollten, oder ob diese nicht eher als eine mögliche Einflussgröße hinsichtlich des klinischen Ergebnisses gelten sollten.

Ein weiterer Kritikpunkt bei diesem Ranking-System ist die Verwendung einer ordinal skalierte Risikostufe, die von höher skalierten Variablen abgeleitet wird. Einerseits wird durch dieses Vorgehen ein Informationsverlust verursacht, andererseits kann dies – insbesondere bei zahlenmäßig kleinen Zentren – zu Verzerrungen führen, beispielsweise wenn zufällig bei der Mehrzahl der Stufenbildungen „aufgerundet“ oder „abgerundet“ wird.

### 1.5.2.2 Das AHRQ Hospital Rating

In jedem Jahr wird von dem Gesundheits-Beratungsunternehmen „Agency for Health Care Research and Quality“ (AHRQ) ein Krankenhaus-Rating veröffentlicht. Die Daten werden aus der USA-weiten Datenbank MedPAR, welches von der „Centers for Medicare and Medicaid Services“ (CMS) bereitgestellt werden, entnommen. Diese Datenbank enthält Daten aus nahezu allen US-amerikanischen Krankenhäusern, wobei Militär- und Veteranenkrankenhäuser ausgenommen sind.

In diesem Rating werden fünf unterschiedliche Dimensionen der Behandlungsqualität betrachtet, für die jeweils einzelne Ratings bzw. entsprechende „Awards“ vergeben werden.

- Für die Patientensicherheit wird für jedes Krankenhaus ein so genannter Sicherheits-Score („Safety Score“) berechnet, der sich aus insgesamt 13 Kenngrößen zusammensetzt. Zu diesen Kenngrößen gehören beispielsweise Tod bei Diagnose mit geringer Mortalität, Dekubitus (Druck-Ulcus), Therapieversagen (keine Heilung), verbliebener Fremdkörper nach Behandlung/Operation, arztbedingter Pneumothorax, durch Behandlung bedingte Infektionen und diverse post-operative Komplikationen. Mittels einer von der öffentlichen „Agency for Healthcare Research and Quality“ bereitgestellten SAS<sup>®</sup>-basierten Software werden so genannte Patient Safety Indicators (PSI) bestimmt.
- Für besonders herausragende klinische Qualität wird ebenso wie für die Patientensicherheit ein jährlicher Preis („Award“) verliehen. Für dessen Bestimmung kommt das folgende mehrstufige Verfahren zum Einsatz:
  - Durchschnitt aller MedPAR Ratings;
  - Sortierung der Krankenhäuser nach ihrem Rating;

- Die besten 20% werden ausgewählt. „Kleine“ Zentren mit weniger als 5.000 Fällen pro Jahr werden ausgeschlossen.
- Die in der Liste verbleibenden Häuser erhalten den Preis.
- Für das Gebiet der Entbindungen und Frauengesundheit werden Daten aus 17 amerikanischen Staaten aus den jeweils vergangenen drei Jahren herangezogen. Bei der Entbindungsqualität spielen die folgenden Faktoren eine Rolle:
  - Kaiserschnitt- oder Dammschnittgröße bei Einzelgeburten;
  - Komplikationsrate bei Kaiser- und Dammschnitten (nur Einzelgeburten);
  - Komplikationsrate bei Kaiserschnitt-Geburten, die die Patientin selbst gewählt hat;
  - Säuglingssterblichkeit, nach Geburtsgewicht stratifiziert.

Schließlich erhält jedes Krankenhaus innerhalb der 17 Staaten ein Rating. Die besten 15% erhalten fünf Sterne, die mittleren 70% erhalten drei Sterne, und die schwächsten 15% erhalten noch einen Stern.

Die Qualität der Frauengesundheit wird anhand der Ergebnisse der Kategorien

- Herz-Bypass,
- Herzklappenersatz,
- andere invasive Eingriffe,
- Herzinfarkt,
- Herzinsuffizienz und
- Schlaganfall

getrennt betrachtet. Die Behandlungsqualität wird risikoadjustiert berechnet, um eine mögliche Heterogenität der Population innerhalb von bestimmten Risikofaktoren zu berücksichtigen. Hier wird ein Rating nach dem oben beschriebenen Verfahren durchgeführt, welches fünf, drei oder einen Stern(e) vergibt.

- Für Sterblichkeit und Komplikationen werden, basierend auf Risikoadjustierungen, Ratings aller in der MedPAR-Datenbank enthaltenen Krankenhäuser berechnet.

- Schließlich wird ein jährlicher Preis für eine besonders herausragende Spezialkompetenz vergeben. Ratings werden für jedes Zentrum innerhalb von therapeutischen Gebieten (wie bspw. Herzinfarkt oder Coronar-Bypass-Behandlung) bestimmt.

Eine detaillierte Methodenbeschreibung findet sich auf der Internet-Seite des Unternehmens [62] bzw. bei Hill et al.[25]

### 1.5.2.3 Krankenhaus-Ranking beim THCIC in Texas

Im Jahre 2002 veröffentlichte das „Texas Health Care Information Council“ (THCIC) in Austin (Texas) ihren Bericht über Qualitätsbeurteilungen von Krankenhauseinrichtungen im gesamten Staat Texas hinsichtlich verschiedener Erkrankungen unter dem Stichwort „Ranking of Hospital Care“ [76].

Als Zielgröße wurden beispielsweise die Mortalitätsraten beim akuten Schlaganfall oder beim akuten Myokardinfarkt betrachtet, die über den gesamten Staat Texas (ca. 34.000 Fälle) gemittelt und mit den betreffenden Zentren verglichen wurden, falls jeweils mindestens fünf 5 Fälle in die Analyse einbezogen werden konnten.

Die Spanne der beobachteten Mortalitätsraten reichte im Jahr 2002 beim akuten Schlaganfall von 4,6% beim St. Anthony Campus in Amarillo (bezogen auf 33 Fälle) bis 13,7% im Hendrick Medical Center in Abilene (314 Fälle).

Auf der Internetseite des THCIC [76] können vollständige Ergebnistabellen zu allen betrachteten Erkrankungen eingesehen werden. Beim akuten Myokardinfarkt liegen die beobachteten (nicht adjustierten) Mortalitätsraten zwischen 0% im Baylor Heart & Vascular Center (Austin) (35 Fälle), im Baylor All Saints Medical Center (30 Fälle), HEALTHSOUTH Rehab Hospital-Wichita (30 Fälle) und im United Regional Health Care System (232 Fälle) bis zu 32,4% im Brownwood Regional Medical Center (34 Fälle).

Für das Ranking der Krankenhäuser wurde die Methodik eingesetzt, die oben zum USWNR-Ranking im Unterpunkt „Outcome“ beschrieben ist. Auf die Bestimmung der ersten beiden Dimensionen (Struktur und Prozess) sowie auf die Berechnung des Mortalitäts-Scores wurde hier verzichtet.

Folgende Limitierungen und Kritikpunkte zu diesem Systems sind festzustellen:

- Die Mortalität wurde ausschließlich als binäre Zielgröße aufgefasst, nicht als „time-to-event“-Variable.
- Die Mortalität wird nur während des Krankenhausaufenthalts betrachtet, dessen Länge im Wesentlichen im Ermessen des Krankenhauses selbst liegt. Die Mortalität wurde nicht auf einen bestimmten Zeitraum bezogen betrachtet (z.B. als 30-Tages-Mortalität).
- In der Veröffentlichung fehlten Angaben zum eingesetzten statistischen Modell für die Risikoklassifizierung sowie zur Modellgüte.
- Für die Erstellung des Konfidenzintervalls der risikoadjustierten Mortalitätsrate fehlte die Berechnungsmethode.

#### **1.5.2.4 Vergleiche zwischen Bypass Chirurgen beim New York State Department of Health**

Das Gesundheitsministerium im Staate New York begann Ende der 1980er Jahre, von allen Bypass-Operationen die Outcome-Daten – und damit verbunden die OP-bedingten Sterbefälle – in einem Register zu sammeln, zu dokumentieren und jährlich zu veröffentlichen. Eine erste Veröffentlichung findet sich bereits 1990, in der die statistisch signifikanten Risikofaktoren für die postoperative Sterblichkeit sowie die entsprechend risikoadjustierte Sterblichkeit nach Krankenhäusern analysiert wurden [16].

Vier Jahre später wurde die Entwicklung der Krankenhaus-Mortalität im Register von 1989 bis 1992 betrachtet. Hierbei zeigte sich eine generelle Abnahme der Sterblichkeitsraten, insbesondere bei denjenigen Kliniken, die zunächst eine hohe Sterblichkeit zeigten [17].

Die Ergebnisse der Analysen wurden aber zusätzlich zur Krankenhausallokation für die einzelnen Chirurgen dargestellt und namentlich veröffentlicht. Dies blieb für die betroffenen Personen nicht ohne Auswirkungen, ungeachtet der statistischen Unsicherheit und der Datenqualität, die den Resultaten zugrunde lag. Diese Vorgänge wurden in einem Artikel der New York Times (Elisabeth Bumiller: „Death-Rate Rankings Shake New York Cardiac Surgeons“) vom 6. September 1995 [68] kritisch diskutiert.

### 1.5.3 Situation in Deutschland

In Deutschland gewinnen Ratings und Rankings unter dem Oberbegriff der „Qualitätssicherung“ zunehmend an Bedeutung. Die Bundesgeschäftsstelle Qualitätssicherung (BQS) gGmbH in Düsseldorf [67], die im September 2000 gegründet wurde und von Gesellschaftern wie der Bundesärztekammer, der Deutschen Krankenhausgesellschaft e.V. sowie von Spitzenverbänden der Krankenkassen geführt wird, versteht sich als unabhängiger Dienstleister auf dem Gebiet der externen vergleichenden Qualitätssicherung im Gesundheitswesen. Die BQS verfolgt mit ihrer Arbeit (etwa 15.000 Krankenhausauswertungen jährlich in etwa 1.700 Krankenhäusern) das Ziel der Darstellung von Qualität im Gesundheitswesen und fühlt sich der Qualitätsverbesserung und damit dem Wohl der Patienten verpflichtet. Die BQS definiert dabei Qualitätsindikatoren und -ziele aus den Bereichen der Diagnosestellung, der Prozesse und der Behandlungsergebnisse.

Darüber hinaus gibt es im Gesundheitswesen weitere vergleichende Systeme und veröffentlichte Analysen, wie etwa

- den „Klinikführer Rhein-Ruhr“, dem etwa 50% der Nordrhein-Westfälischen Krankenhäuser angehören (federführend ist der Initiativkreis Ruhrgebiet) und der einmal jährlich als gebundenes Buch erscheint [28] bzw. aus dem Internet herunterladbar ist [71],
- den „Klinikführer Berlin“ (mit einer Teilnahmequote von über 90%), der von der Tageszeitung „Der Tagesspiegel“ durchgeführt wird und ebenfalls herunterladbar ist [75], oder
- den aktuellen „Bericht zur Qualitätssicherung der stationären Versorgung mit Routinedaten (QSR)“ des AOK-Bundesverbands [2], bei dem die Behandlungsqualität für verschiedene Indikationen – so auch dem akuten Myokardinfarkt – basierend auf risiko-adjustierten linearen Modellen ausgewertet wurde.

Rankings werden natürlich nicht nur über medizinische Versorgungseinrichtungen, sondern häufig auch über Ausbildungseinrichtungen angestellt. Neben dem allgemein bekannten PISA-Test, bei dem die Leistungen von Schülern über die OSZE-Staaten wiederholt miteinander verglichen werden, gab es auch in Deutschland Vergleiche zwischen Universitäten, wie beispielsweise die Abschlußnoten von Absolventen der

medizinischen Fakultäten [6], was hier nur als ein Beispiel von vielen herausgegriffen wird.

## 1.6 Ranking-Systeme in der Diskussion zur Öffentlichen Gesundheit

### 1.6.1 Interviews mit Gesundheitsexperten

Im Frühjahr des Jahres 2004 fand eine Gesprächsrunde zur generellen Bedeutung von Krankenhaus-Rankings für die „Kunden“ bzw. für das Gesundheitssystem insgesamt statt, bei der die US-amerikanischen Gesundheitsexperten Jim Blazar, Martyn Howgill, Ken Foster und Lynne Cunningham befragt wurden [27].

Eine Umfrage aus dem Jahr 2003, durchgeführt vom *Wall Street Journal*, ergab, dass 26% der Bevölkerung Krankenhaus-Rankings als eines der wichtigsten Indikatoren für die medizinische Versorgungsqualität ansehen. 10% der Befragten gaben an, sie würden die Ergebnisse dieser Krankenhaus-Evaluationen als Entscheidungskriterium für die Wahl eines Hospitals benutzen. Die Wichtigkeit für die Bevölkerung scheint zugenommen zu haben, da frühere Studien eher geringeres Interesse an Krankenhaus-Rankings gezeigt hatten.

Die folgenden Positionen hinsichtlich der Bedeutung von Rating-Systemen für die Verbraucher wurden vertreten:

- Martyn Howgill, Vize-Präsident für internationale Geschäftsentwicklung im M.D. Anderson Krebszentrum der Universität von Texas in Houston, gibt an, dass zwar einige seiner Patienten aufgrund des Top-Rankings in seine Klinik kamen. Andererseits glaubt er nicht an eine große Zahl dieser „self-selecting patients“, da viele Patienten aufgrund ihrer geographischen Lage oder mangelnder Mobilität einerseits bzw. aufgrund ihrer Erkrankungsart (z.B. Notfallindikationen) andererseits kaum eine wirkliche Auswahlmöglichkeit besitzen. Die Wichtigkeit und Relevanz der Rankings wird von ihm zwar anerkannt, jedoch schätzt er deren Bedeutung für die Versorgungsqualität eher skeptisch ein.

- Jim Blazar, Marketing-Leiter der Cleveland Clinic Foundation, hingegen sieht einen Trend hin zu mehr Interesse an Transparenz der Qualität der Leistungserbringer. Aufgrund der zunehmenden Fülle von derartigen Rating-Systemen wird das Informationsangebot seiner Ansicht nach für die amerikanischen Verbraucher zunehmend unübersichtlich und schwierig, die für sie gewünschten Informationen zu erlangen.
- Kenneth L. Foster, derzeit als Vize-Präsident für regionale Entwicklung und strategische Planung bei BryanLGH Health System in Lincoln (Nebraska) tätig, sieht zur Zeit nur die Spitze eines Eisbergs bei Rating-Systemen. Seiner Ansicht nach sind jedoch die Informationsquellen und die Fähigkeit der Verbraucher, sie zu nutzen, noch nicht vollständig entwickelt.
- Für Lynne Cunningham, Vorsitzende der eigenen Beratungsagentur für strategische Planung und qualitative Marktforschung, ist die Veröffentlichung der Ratings für den Verbraucher weitgehend bedeutungslos. Laut einer Studie des *Wall Street Journal* werden die Ranking-Ergebnisse von weniger als einem Prozent der Verbraucher, die diese zur Kenntnis erhalten, auch tatsächlich für ihre Wahl des Krankenhauses benutzt.

### 1.6.2 Beitrag aus der Beratungswirtschaft

Julia A. Rieve [46], seit 25 Jahren auf dem Gebiet des klinischen Managements tätig und Gründerin und Präsidentin eines Health Care Beratungsunternehmens, äußert sich zur US-nationalen Bedeutung von Krankenhaus-Rankings aus der Perspektive des „Case-Managers“.

Sie führt einen großen Teil der steigenden Kosten im Gesundheitswesen auf Behandlungs- oder Prozessfehler zurück. Obwohl die direkte monetäre Zuordnung der Fehler schwierig ist und von Betrachter zu Betrachter variiert, ist von jährlichen Beträgen in \$-Milliardenhöhe auszugehen. Daher beurteilt sie die Bedeutung von Krankenhaus-Rankings als essenziell für die Verbesserung der Behandlungsqualität und sieht dabei den Case Manager in einer Schlüsselrolle, indem er die Ergebnisse von Krankenhaus-Rankings beurteilt und in seine Behandlungsentscheidung einfließen lässt.

## 1.7 Ranking-Systeme aus statistischer Sicht

In einem kritischen Artikel nehmen Lars A. Endahl und Jan Utzon [10] Stellung zu den Projekten, die im Dänischen Gesundheitswesen zur Beurteilung von Qualitätsindikatoren diskutiert werden.<sup>1 2</sup>

Die Autoren sehen die uneingeschränkte Veröffentlichung von Qualitätsindikatoren kritisch, da die Darstellung der Ergebnisse immer ein vereinfachtes Bild der Wirklichkeit zeigt. Ein Krankenhaus mit geringer Qualität in einem untersuchten Bereich könnte auf anderen Gebieten durchaus hervorragende Qualität aufweisen, die aber im Rahmen der Beurteilungsstudie nicht untersucht wurden bzw. gar nicht aufzudecken sind. Bei der Einführung von (Risiko-)Adjustierungen besteht die Gefahr, dass diese möglicherweise nicht der Fragestellung entsprechend bzw. fehlerhaft definiert sind, und daher zu falschen Ergebnissen führen können.

Weitere Schlüsselprobleme für Krankenhaus-Ranglisten sind aus Sicht der Autoren:

- Der Datenbestand ist häufig zu gering, um „scharfe“ Aussagen über die tatsächliche Rangfolge zwischen den Krankenhäusern treffen zu können. Falls die Rangfolge auf Basis der Punktschätzer aufgestellt wird, werden sich die Konfidenzbereiche eines Indikators bei kleinen Fallzahlen pro Zentrum stark überlappen.

*Eigene Anmerkung:*

*Alternativ könnte die Rangfolge auch mittels der unteren Konfidenzgrenzen bestimmt werden; diese ist jedoch stark von der zugrunde liegenden Fallzahl abhängig. Ein „kleines“ gutes Zentrum würde hier schlechter abschneiden als ein weniger gutes Zentrum mit höherer Patientenzahl.*

- Aufgrund der unterschiedlichen Patientenzahlen ergibt sich weiterhin das Problem, dass sich aufgrund der zufälligen Streuung bei gleicher Behandlungsqualität „kleinere“ Zentren eher am oberen bzw. unteren Ende der Reihenfolge befinden als „größere“ Zentren.

---

<sup>1</sup>„Das Nationale Indikatorprojekt“

<sup>2</sup>„Zentrum zur Evaluierung und Medizinischen Technologieabschätzung“

- Bei Betrachtungen über längere Zeitspannen hinweg bzw. bei wiederholten Analysen besteht zusätzlich der Effekt, dass sich zunächst gute und schlechte Zentren bei späteren Analysen eher in Richtung der Mitte orientieren werden, selbst bei konstanter Behandlungsqualität („Regression towards the mean“).
- Schließlich kritisieren die Autoren, dass bei Ranglisten keine klinische Definition für gute, zufrieden stellende bzw. nicht zufrieden stellende Behandlungsqualität gegeben ist. Dies hat zur Folge, dass möglicherweise – zu Unrecht – zufrieden stellend behandelnde Zentren als ungenügend, oder umgekehrt nicht zufrieden stellende Zentren als gut angesehen werden, da der Ranglistenplatz von der Qualität in der Stichprobe und vom Gesamtniveau abhängt.

*Eigene Anmerkung: Dies ist beispielsweise bei dem weiter oben beschriebenen Rankingsystem der „America’s Best Hospitals“ in den USA der Fall.*

Folgende Anforderungen an Qualitätsindikatoren werden von den Autoren gestellt:

- Validität: Der Untersuchungsgegenstand muss durch die Messung tatsächlich abgebildet werden.
- Zuverlässigkeit/Reproduzierbarkeit: Die Ergebnisse sollen nicht durch (zufällige) Messfehler beeinflusst sein.
- Die Indikatoren sollen risikoadjustiert werden, damit nicht die Krankenhäuser, die Patienten mit a priori schlechterer Prognose aufnehmen, schlechter beurteilt werden.
- Die Risikoadjustierung muss auf guter Datenqualität basieren. Unter Umständen sind die wirklichen Risikofaktoren nur lückenhaft erfasst bzw. gar nicht erhoben worden (wie z.B. ungesunde Lebensweise). Gegebenenfalls sollen die Registerdaten um zusätzliche Variablen erweitert werden (sofern dies möglich ist).
- Weiterhin sollen die Risikoadjustierungen möglichst einfach interpretierbar sein. Das heißt, dass auch die Ursachen für mögliche Unterschiede in der Qualität zwischen den Zentren bzw. auch im internationalen Vergleich aus den Daten ersichtlich sind.

- Eine weitere Bedingung für die Evaluierung von Qualitätsindikatoren ist die Relevanz für die Öffentlichkeit. Nur wenn ein hinreichendes öffentliches Interesse an der Problematik besteht, werden die Ergebnisse von der Öffentlichkeit im Sinne der Krankenhausauswahl genutzt werden.
- Nur wenn die Ergebnisse adäquat und verständlich dargestellt werden, können diese für Entscheidungsprozesse in der Klinik genutzt werden.

Auf der einen Seite sind (eindimensionale) Ranglisten für die Öffentlichkeit leicht interpretierbar, auf der anderen Seite jedoch verbleiben viele offene Fragen, die mit einer solchen Darstellung nicht beantwortet sind, wie etwa:

- die Frage, mit welcher Berechnungsmethode die Indikatoren bestimmt wurden,
- auf welche Messungen sich die Ergebnisse beziehen, oder
- mit welcher Unsicherheit ein bestimmter Qualitätsindikator – und damit der entsprechende Ranglistenplatz des Krankenhauses – belegt ist. So ist z.B. nicht unerheblich zu wissen, ob ein Krankenhaus, das auf der Liste auf Platz 10 (von 50) liegt, mit einer Sicherheit von 95% zwischen Platz 8 und 12 oder zwischen Platz 2 und 35 liegt.

*Eigene Anmerkung:*

*Als Vorschlag für eine Strategie kann man zunächst nach Unterschieden zwischen den Häusern generell fragen (F-Test oder ähnlicher Test über eine Globale Hypothese). Nur wenn die Frage – mit geeignetem Fehlerrisiko abgesichert – mit „ja“ beantwortet werden kann, sollte untersucht werden, zwischen **welchen** Zentren die Unterschiede bestehen. Auf Basis von Vertrauensbereichen kann dann eine Klassifikation im Sinne von „besser als der Durchschnitt“, „nicht verschieden vom Durchschnitt“, „schlechter als der Durchschnitt“ durchgeführt werden. Verschiedene Möglichkeiten, wie paarweise Vergleiche, Vergleiche gegen eine Benchmark, Vergleiche gegen den Mittelwert oder auch „many-to-one“-Vergleiche (ist ein bestimmtes Zentrum signifikant besser/schlechter als der Rest), stehen zur Auswahl.*

*Die Frage nach der Signifikanz – die mit der Breite des Konfidenzintervalls zusammenhängt – hängt nicht nur vom Erwartungswert des Zentrums, sondern auch von der Fallzahl a) des gesamten Versuchs und b) der beim Vergleich betrachteten Zentren ab. Ein in Wahrheit „schlechtes“ Zentrum hat bei kleinerer Fallzahl bessere Chancen, als solches unentdeckt zu bleiben als mit größerer.*

## 1.8 Diskussion der Literatur und Schlussfolgerung

Zum gegenwärtigen Zeitpunkt sind Beurteilungsstudien oder Beurteilungssysteme von Versorgungseinrichtungen in den Vereinigten Staaten wesentlich weiter verbreitet als in Kontinental-Europa (d.h. ohne Großbritannien). Dies zeigt auch die Zahl der Veröffentlichungen in diesem Bereich, die zu einem sehr großen Anteil aus den USA stammen bzw. auf US-amerikanischen Ergebnissen beruhen. Ein wichtiger Grund hierfür dürfte in der im Vergleich zu Deutschland und Europa deutlich stärker ausgeprägten Wettbewerbssituation liegen, in der die dortigen Leistungserbringer zueinander stehen. Zudem ist die Gesetzgebung zum Datenschutz in den USA weniger strikt als in Europa.

Trotzdem ist der Nutzen für den Verbraucher durchaus umstritten, insbesondere da es an Einheitlichkeit der Systeme und an der Nutzbarkeit im praktischen Einzelfall mangelt.

Weiterhin sind Krankenhaus-Rankings aufgrund von oft schwacher Datenlage, Zufallseffekten, zufälligen Qualitätsschwankungen über die Zeit und mangelnder Interpretierbarkeit der Ergebnisse hinsichtlich der Qualität aus statistischer Sicht umstritten. Bei der Darstellung, Veröffentlichung und Interpretation von Ranglisten besteht daher die Gefahr, dass die Ergebnisse in der Öffentlichkeit stark vereinfacht wahrgenommen werden (z.B. nur die Punktschätzer werden verstanden und betrachtet), obwohl die Situation komplexer ist.

---

## 2 Statistische Methodik

### 2.1 Einführung

In diesem Kapitel werden statistische Analysemethoden vorgestellt, die im Rahmen dieser Arbeit diskutiert werden. Hierbei wird kein Anspruch auf Vollständigkeit aller verfügbaren Methoden erhoben. Dieses Kapitel soll vielmehr als Einführung in die Methoden verstanden werden, die bei der Zielstellung dieser Arbeit von Bedeutung sind.

Bei Vergleichen von medizinischen Einrichtungen bzw. Kliniken auf Basis von Registerdaten – die im Folgenden als „Einrichtungsvergleiche“ bezeichnet werden – sind zur Planung der Analyse zunächst verschiedene Fragen zu beantworten:

- **Anzahl der Zielgrößen:**

Zum einen gibt es Situationen, in denen die Behandlungsqualität (also der Behandlungserfolg) anhand *eines* Indikators gemessen werden soll. In diesem Fall spricht man von einem RANKING. Werden aber mehrere Zielgrößen betrachtet, die innerhalb eines Versuchs / einer Auswertung dargestellt werden sollen, spricht man von einem PROFILING. Beim Profiling werden also mehrere Rankings gleichzeitig durchgeführt, bei denen die Zentren im allgemeinen verschiedene Rangplätze einnehmen. So kann untersucht werden, ob sich in den Rängen bestimmte Strukturen zeigen, welches die Stärken und Schwächen der Einrichtungen sind oder zu welchem „Preis“ ein gutes Abschneiden in einem Parameter erzielt wird.

Im Rahmen dieser Arbeit wird hauptsächlich auf die methodischen Aspekte der Rankings eingegangen; das Profiling von Krankenhäusern ist für die Methodendiskussion weniger relevant, da es sich hierbei im Wesentlichen um mehrfaches Ranking handelt, und daher nur am Rande diskutiert wird.

- **Skalenniveau der Zielgröße(n):**

In den meisten Untersuchungen an Registerdaten liegen die Zielgrößen als stetige und als normalverteilt angenommene Variablen oder in binärer Form vor. Es sind aber Situationen vorstellbar, bei denen die Ausprägungen in geordneten Stufen (Ordinalskala) oder als Zeit bis zu einem Ereignis („time-to-event“) gemessen werden. In seltenen Fällen kann eine nominal skalierte Zielgröße vorliegen, bei der also keine Rangfolge zwischen den Ausprägungen existiert. Zusätzlich zu den klassischen Skalenniveaus könnte eine Erhebung im Gesundheitswesen auch Daten, die einen (stochastischen) Prozess beschreiben, betrachten.

Für diese Arbeit werden aufgrund der Relevanz im Wesentlichen die beiden zuerst genannten Skalenniveaus diskutiert. Für die anderen Situationen werden an geeigneter Stelle Ausblicke gegeben.

- **Einflussgrößen und Adjustierung:**

Wie bereits im ersten Kapitel diskutiert wurde, ist es bei Einrichtungsvergleichen zwingend erforderlich, die Ergebnisse hinsichtlich der Risikofaktoren, die für den Behandlungs(miss)erfolg bedeutsam sind und die sich zumeist zwischen den Kliniken unterscheiden, zu adjustieren. Diese Faktoren müssen – zusätzlich zur Einflussgröße *Klinik* – in das finale Modell einbezogen werden. Zur Identifikation der relevanten Faktoren sollten Substanzwissenschaftler (z.B. Fachexperten) hinzugezogen werden. Ist dies nicht möglich, muss auf Basis von signifikanten Faktoren bzw. Faktoren mit starkem Einflussgrad (Steigungsparameter/slopes oder Odds Ratios) gearbeitet werden. Die Auswahl der Faktoren ist in der Regel nicht eindeutig zu klären; man sollte aber bei hinreichend großer Fallzahl (und damit hoher Zahl von Freiheitsgraden) nicht zu „sparsam“ mit den Adjustierungsfaktoren umgehen. Eine Modellüberspezifizierung hätte einen geringeren Einfluss auf das Ergebnis als eine Unterspezifikation.

- **Hierarchie der Datenlage (Cluster-Struktur):**

Wie ebenfalls im ersten Kapitel erwähnt wurde, sollte bei Einrichtungsvergleichen die Korrelation der Beobachtungen *innerhalb* der Einrichtungen Berücksichtigung finden. Registerdaten sind in der Terminologie der Stichprobenverfahren also als Klumpenstichproben zu begreifen, da Patienten praktisch immer innerhalb genau eines Zentrums erfasst werden. Datenlagen dieser Form nennt man „hierarchische Daten“ oder auch „multilevel data“ oder „clustered data“.

Die beobachteten Daten von Patienten können somit als Messwiedholungen desselben Zentrums aufgefasst werden; in diesem Fall spricht man bei den Patienten von „level-1“ und bei den Kliniken von „level-2“-Daten. Die hierarchische Situation kann auf mehr als zwei Ebenen erweitert werden, etwa wenn bei den Patienten mehrere Messungen durchgeführt werden. Dann sind die Messungen innerhalb des Patienten wiederum untereinander korreliert. Diese – für die Analyse wegen der zu modellierenden (und häufig unbekannt) Kovarianz-Struktur wesentlich komplexere und schwieriger zu modellierende – Situation ist bei Einrichtungsvergleichen jedoch seltener, da wiederholte Messungen häufig in wiederholten Rankings auftreten.

- **Modellwahl:**

Schließlich muss entschieden werden, mittels welcher statistischen Modellklasse die Auswertung geschehen soll. Es bieten sich – je nach dem Skalenniveau der Zielgröße – neben den klassischen linearen Modellen (Varianzanalyse, Regression, Kovarianzanalyse) einige alternative Verfahren an, die im Folgenden diskutiert werden. Bei der Wahl des Modells ist durch die hierarchische Datenlage allerdings ein limitierender Faktor gegeben, da diese von vielen Analysemethoden nicht berücksichtigt werden kann. Lineare Modelle können jedoch sowohl für hierarchische Daten als auch für gemischte (feste und zufällige) Effekte angewendet werden. Eine Auswahl der zur Verfügung stehenden Modelle wird in diesem Kapitel vorgestellt und diskutiert.

## Statistische Aufgabenstellung

Ziel dieser Arbeit ist es, Eigenschaften von statistischen Methoden allgemein und für das Anwendungsbeispiel speziell aufzubereiten und im Sinne eines Analysekonzeptes zu bewerten.

Für das Analyseergebnis dieser Arbeit sind zwei Zielstellungen zu betrachten:

- **Methodenteil:** Identifizierung geeigneter statistischer Methoden für die Analyse von Daten aus der Versorgungsforschung (Registerdaten);
- **Analyseteil:** Ermittlung optimaler Behandlungsstandards für Patienten (oder Patientengruppen).

In diesem Kapitel (Methodenteil) wird auf der Basis der vorhandenen Modellierungsansätze ein Analysekonzept mit dem Ziel der Benchmarkbildung in der Versorgungsforschung erarbeitet.

Hierzu ist zunächst die Einführung des klassischen linearen Modells hilfreich.

**Definition – Das lineare Modell:**

Seien

$y, e$  zufällige Vektoren mit je  $N$  Komponenten;

$\beta$  fester Vektor mit  $k$  Komponenten;

$X$  Matrix der Dimension  $Nk$ .

Ein LINEARES MODELL ist eine Darstellung der Form

$y = X\beta + e,$  wobei

$y = (y_1, \dots, y_N)'$  Realisationen der abhängigen Variablen  $Y$ ;

$X$  Designmatrix;

$\beta = (\beta_1, \dots, \beta_k)$  Parametervektor;

$e = (e_1, \dots, e_N)'$  Zufallsfehler mit unabhängig identisch verteilten Komponenten

$E(e) = 0$ ;  $\Sigma_e = \sigma_e^2 V$  und  $\sigma_e^2 > 0$

Der Wert  $N$  entspricht der Gesamtzahl der als untereinander unabhängig angenommenen Beobachtungen.

Eine Lösung  $\hat{\beta} \in \mathbb{R}^k$  des Normalgleichungssystems  $X'X\hat{\beta} = X'y$  bezeichnet man auch als **Kleinste-Quadrate-Schätzung** für  $\beta$ . Ist  $\beta$  linear und erwartungstreu schätzbar, so gilt nach dem Gauß-Markov-Theorem die Existenz und die Eindeutigkeit dieser Lösung  $\hat{\beta}$  („Gauß-Markov-Schätzer“).

Auf weitere theoretische Grundlagen der Linearen Modelle wird im Rahmen dieser Arbeit nicht weiter eingegangen. Hierzu sei auf die einschlägigen Lehrbücher (z.B. von Scheffé [49] oder Hartung [22]) verwiesen.

Im Folgenden werden die gängigsten Typen von linearen Modellen eingeführt, die für die Analyse von Daten in der Versorgungsforschung von möglicher Relevanz sind, d.h. in welcher Weise die eingangs benannten Fragestellungen und hinsichtlich der Situation bei Einrichtungsvergleichen modelliert werden können. Die Einführung der Methode und Notation erfolgt schrittweise, jedoch ohne auf maß- und wahrscheinlichkeitstheoretische Grundlagen zurückzugreifen.

## 2.2 Klassische Lineare Modelle

Für dieses Kapitel gilt die folgende Nomenklatur:

$Y$	Zielgröße
$y$	Ausprägung der Zielgröße $Y$
$y_{ij}$	Ausprägung der Zielgröße $Y$ in Stufe $i$ für die Messwiederholung $j$
$\bar{y}_i$	Mittelwert der Zielgröße $Y$ in der $i$ -ten Stufe
$h_i$	beobachtete relative Ereignishäufigkeit in der $i$ -ten Stufe
$X$	Designmatrix der festen Effekte
$X_i$	$i$ -te feste Einflussgröße ( $i = 1, \dots, k$ )
$\beta$	Parametervektor für feste Effekte (mit $\beta = (\beta_1, \dots, \beta_k)'$ )
$\mu$	Populationsmittel
$\tilde{p}$	Grundwahrscheinlichkeit für die Population
$\mu_i$	Erwartungswert der $i$ -ten Stufe
$p_i$	Ereigniswahrscheinlichkeit der $i$ -ten Stufe
$Z$	Designmatrix der zufälligen Effekte
$Z_i$	$i$ -te Einflussgröße mit zufälligen Effekten
$\alpha$	Irrtumswahrscheinlichkeit (Fehler 1. Art)
$\gamma$	Parametervektor für zufällige Effekte
$e$	zufälliger Restfehler
$E()$	Erwartungswert
$k_o$	obere Grenze des Konfidenzintervalls
$k_u$	untere Grenze des Konfidenzintervalls
$k$	Anzahl der modellierten festen Einflussgrößen $X_i$
$p$	Anzahl der Ausprägungen einer kategoriellen Einflussgröße $X_i$ bzw. $Z_i$
$q$	Wiederholungszahl innerhalb der Einrichtungen (balancierter Fall)
$n_i$	Wiederholungszahl innerhalb der $i$ -ten Einrichtung (unbalancierter Fall)
$N$	Gesamtfallzahl des Versuchs
$G$	Kovarianzmatrix der zufälligen Effekte $\gamma$ im gemischten Modell
$R$	Kovarianzmatrix der Restfehler $e$ im gemischten Modell
$C$	Kovarianzmatrix der Schätzfehler ( $\hat{\beta} - \beta, \hat{\gamma} - \gamma$ ) im gemischten Modell
$V$	Kovarianzmatrix aller Beobachtungen

### 2.2.1 Varianzanalyse

Bei der Varianzanalyse (zurückgehend auf Sir Ronald Aylmer Fisher, 1890-1962) wird der (mögliche) Einfluss von mehreren gruppiert vorliegenden Faktoren auf den Untersuchungsgegenstand (Zielgröße) gemessen. Die Einflussfaktoren können dabei von den Versuchsobjekten (z.B. Patienten) selbst getragen werden (wie etwa das Geschlecht) oder diesen erst zugewiesen werden (etwa eine Behandlung).

Zunächst untersucht man, ob *überhaupt* ein Einfluss vorliegt und, wenn ja, dann *in welcher Stärke* der Einfluss besteht. Die Varianzanalyse beantwortet also die Frage, ob die verschiedenen Stufen eines Einflussfaktors *statistisch signifikant* unterschiedliche Wirkungen auf die Zielgröße besitzen und sie quantifiziert ihre Effekte.

#### 2.2.1.1 Einfache Varianzanalyse, Einfachklassifikation

Bei der einfachen Varianzanalyse (auch „One-Way ANOVA (Analysis of Variance)“ oder „Einfachklassifikation genannt) wird ein Einflussfaktor (also  $k = 1$ ) mit  $p$  Ausprägungen (Stufen) im Modell betrachtet. Dabei wird zunächst davon ausgegangen, dass es sich bei der Untersuchung der  $p$  **unabhängigen** Messreihen um (vollständig) randomisierte Versuchspläne handelt, d.h. dass die Versuchsobjekte den Stufen der Einflussgröße zufällig und nicht systematisch zugewiesen werden. Ist dies nicht der Fall, ist die Anwendung dieses Modells nicht sinnvoll (siehe Diskussion am Ende dieses Abschnitts).

Für die Elemente der Designmatrix  $X$  gilt:

$$x_{ij} \in \{0, 1\} \quad \text{für} \quad i = 1, \dots, p \quad \text{und} \quad j = 1, \dots, n_i .$$

#### Beispiel

Es sollen die Einflüsse von  $p$  Kliniken auf den mittleren Blutdruck bei Entlassung von Hypertonie-Patienten untersucht werden. Der Blutdruck von  $N = n_1 + n_2 + \dots + n_p$  Patienten ist dann im Vektor  $y = (y_{ij})$  (mit  $i = 1, \dots, p$  und  $j = 1, \dots, n_i$ ) realisiert.

Das Modell lässt sich dann in Matrixschreibweise darstellen als

$$y = X\beta + e = \begin{pmatrix} y_{1_1} \\ y_{1_2} \\ \vdots \\ y_{1_{n_1}} \\ y_{2_1} \\ y_{2_2} \\ \vdots \\ y_{2_{n_2}} \\ \vdots \\ \vdots \\ y_{p_1} \\ y_{p_2} \\ \vdots \\ y_{p_{n_p}} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ 1 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \\ 0 & 0 & 0 & \cdots & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \vdots \\ \beta_p \end{pmatrix} + \begin{pmatrix} e_{1_1} \\ e_{1_2} \\ \vdots \\ e_{1_{n_1}} \\ e_{2_1} \\ e_{2_2} \\ \vdots \\ e_{2_{n_2}} \\ \vdots \\ \vdots \\ e_{p_1} \\ e_{p_2} \\ \vdots \\ e_{p_{n_p}} \end{pmatrix}$$

In der Regel wird zur Modellspezifikation nicht die Matrixschreibweise, sondern die Schreibweise mit den Faktoren als Summanden benutzt. Für das beschriebene Beispiel schreibt sich das Modell dann:

$$y_{ij} = \beta_i + e_{ij} \quad \text{mit} \quad i = 1, \dots, p \quad \text{und} \quad j = 1, \dots, n_i .$$

Der Parametervektor  $\beta$  wird häufig dargestellt als  $\beta = \mu + \alpha$ , wobei  $\mu$  als (invariantes) Absolutglied und  $\alpha = (\alpha_1, \alpha_2, \dots, \alpha_p)'$  als die Gruppeneffekte aufgefasst werden.

Das Modell schreibt sich dann zu

$$y_{ij} = \mu + \alpha_i + e_{ij} \quad \text{mit} \quad i = 1, \dots, p \quad \text{und} \quad j = 1, \dots, n_i ,$$

mit  $\mu$  als Gesamtmittelwert und die  $\alpha_i$  mit  $\sum_{i=1}^p \alpha_i = 0$  als Gruppeneffekten.

Die Designmatrix  $X$  wird hierzu um die Spalte  $(1, \dots, 1)' = \mathbf{1}_n$  erweitert. Der Parametervektor  $\beta$  schreibt sich somit zu

$$\beta = (\mu, \alpha_1, \alpha_2, \dots, \alpha_p)' .$$

Für die Gesamtfallzahl des Experiments gilt stets  $N = \sum_{i=1}^p n_i$ , wobei  $N > p$  sein muss.

### Punktschätzung in der einfachen Varianzanalyse

Die gemessenen Gruppenmittelwerte  $\bar{y}_i$  dienen im Modell als Schätzer für die  $\beta_i$ :

$$\begin{aligned}\hat{\beta}_i &= b_i = \bar{y}_i && \text{bzw.} \\ \hat{\mu} &= m = \bar{y}_{..} && \text{und} \\ \hat{\alpha}_i &= a_i = \bar{y}_i - \bar{y}_{..} && ,\end{aligned}$$

mit

$$\begin{aligned}\bar{y}_i &= \frac{1}{n_i} \sum_{j=1}^{n_i} y_{ij} && \text{und} \\ \bar{y}_{..} &= \frac{1}{n} \sum_{i=1}^p \sum_{j=1}^r y_{ij} && .\end{aligned}$$

Für die Kovarianzmatrix des Restfehlers gilt hier bei angenommener Homoskedastizität:

$$\Sigma_e = \Sigma_y = \sigma^2 I_N .$$

### Hypothesentests in der einfachen Varianzanalyse

Die Analysetafel für die Einfachklassifikation hat die in Tabelle 2.1 dargestellte Gestalt.

Die Abkürzungen der Tabelle gehen ihrerseits zurück auf Fisher [13] und bezeichnen:

SS	<u>S</u> um of <u>S</u> quares (Quadratsumme allgemein)
SST	<u>S</u> um of <u>S</u> quares for <u>T</u> reatments (Quadratsumme der Behandlungen)
SSE	<u>S</u> um of <u>S</u> quares for <u>E</u> rrors (Fehlerquadratsumme)
SSG	<u>G</u> rand <u>S</u> um of <u>S</u> quares (Gesamtquadratsumme)
MS	<u>M</u> ean <u>S</u> quare (Mittlere Quadratsumme allgemein)
MST, MSE	<i>entsprechend.</i>

Zur Prüfung, ob tatsächlich signifikante Unterschiede zwischen den  $p$  Behandlungsstufen bestehen, wird folgender Hypothesensatz zugrunde gelegt.

$$\begin{aligned}H_0 : \quad & \beta_i = \beta_j && \text{für alle } i \neq j \in (1, \dots, p) \quad \text{vs.} \\ H_1 : \quad & \beta_i \neq \beta_j && \text{für wenigstens ein } i \neq j \in (1, \dots, p),\end{aligned}$$

**Tabelle 2.1: Tafel der einfachen Varianzanalyse (ANOVA)**

Streuungs- ursache	Freiheits- grade	Quadratsummen	Mittlere Quadratsumme
Unterschiede zwischen den Messreihen	$p - 1$	$SST = \sum_{i=1}^p n_i (\bar{y}_i - \bar{y}_{..})^2$	$MST = \frac{SST}{p - 1}$
zufälliger Fehler	$N - p$	$SSE = \sum_{i=1}^p \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2$	$MSE = \frac{SSE}{N - p}$
Gesamt	$N - 1$	$SSG = \sum_{i=1}^p \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{..})^2$	

mit  $\mu_i$ : wahrer unbekannter Mittelwert in Gruppe  $i$ .

Bezogen auf die Modelldarstellung mit Gesamtmittelwert  $\mu$  und Gruppeneffekten  $\alpha_i$  wird folgende Schreibweise des Hypothesensatzes verwendet:

$$\begin{aligned}
 H_0 : \quad & \alpha_i = 0 && \text{für alle } i \in (1, \dots, p) && \text{vs.} \\
 H_1 : \quad & \alpha_i \neq 0 && \text{für wenigstens ein } i \in (1, \dots, p),
 \end{aligned}$$

mit  $\alpha_i$ : wahrer unbekannter Effekt in Gruppe  $i$ .

$H_0$  heißt „Nullhypothese“,  $H_1$  wird als „Gegenhypothese“ oder auch „Alternativhypothese“ bezeichnet.

Zur Prüfung von  $H_0$  werden die Terme aus obiger Tafel betrachtet. Da unter  $H_0$

$$E_{H_0}(\hat{\alpha}_i) = \mu$$

gilt, kann nämlich bei hinreichend großer Streuung zwischen den Behandlungen im Verhältnis zu der Restfehlerstreuung (auch „Residualstreuung“ genannt) auf das Vorliegen von  $H_1$  geschlossen werden.

$H_0$  soll nun mit vorgegebener Irrtumswahrscheinlichkeit („Fehler 1. Art“)  $\alpha$  (häufig:  $\alpha = 5\%$ ) geprüft werden. Dazu wird die Prüfgröße (Teststatistik)  $F$  gebildet aus den

Werten MST und MSE aus obiger Tabelle und kann angenommen werden als unter der Nullhypothese  $F$ -verteilt mit  $p - 1$  und  $N - p$  Freiheitsgraden:

$$F = \frac{\text{MST}}{\text{MSE}} \sim F_{p-1, N-p} .$$

Überschreitet der Wert der Teststatistik  $F$  eine kritische Grenze  $c_\alpha$ , mit

$$c_\alpha = F_{p-1, N-p, 1-\alpha} ,$$

wird  $H_0$  zugunsten von  $H_1$  abgelehnt, d.h. es wird auf signifikante Unterschiede zwischen den Gruppen geschlossen.

Ist die Anzahl der Gruppen (bzw. Stufen)  $p$  größer als 2, kann man allerdings zunächst nicht bestimmen, zwischen *welchen* Gruppen die Unterschiede bestehen. Hierzu sei auf *multiple Vergleiche* oder auch *geschlossene Testprozeduren* verwiesen (siehe etwa Bauer et al. [3]).

### **Intervallschätzung in der einfachen Varianzanalyse**

Nach Bestimmung der Punktschätzer und der Testentscheidungen ist von Interesse, in welchem Bereich die Gruppenmittelwerte bzw. meist die Unterschiede zwischen den Gruppen – mit vorgegebener Sicherheitswahrscheinlichkeit (z.B. 95%) – liegen.

Für die Beschreibung der Schätzgenauigkeit der beobachteten Gruppenmittelwerte gilt:

**Satz: Wurzel-n-Gesetz:**

Sind die Zufallsvariablen  $y_{i1}, \dots, y_{in_i}$  identisch verteilt mit Erwartungswert  $\beta_i$ , so gilt

$$E(\bar{y}_i) = \beta_i .$$

Sind zusätzlich  $y_{i1}, \dots, y_{in_i}$  unabhängig und ist ihre Varianz gleich  $\sigma_i^2$ , so gilt zusätzlich

$$\text{Var}(\bar{y}_i) = \frac{\sigma_i^2}{n_i} \quad \text{bzw.} \quad \sigma(\bar{y}_i) = \frac{\sigma_i}{\sqrt{n_i}} .$$

Unter diesen Voraussetzungen ist  $\bar{y}_i$  asymptotisch normalverteilt.  $\sigma(\bar{y}_i)$  heißt Standardfehler (Standard error of the mean, SEM) des Gruppenmittelwerts  $\bar{y}_i$  mit

$$\hat{\sigma}_i = \sqrt{\frac{\sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2}{n_i - 1}} .$$

Somit können für die wahren Gruppenmittelwerte  $\beta_i$  unter Verwendung des mit  $\hat{\sigma}_i$  geschätzten Standardfehlers (Varianz unbekannt) Vertrauensintervalle mit Abdeckungssicherheit  $1 - \alpha$  der folgenden Form angegeben werden:

$$P \left( \beta_i \in \left[ \bar{y}_i - t_{r;1-\alpha/2} * SEM(\bar{y}_i) ; \bar{y}_i + t_{r;1-\alpha/2} * SEM(\bar{y}_i) \right] \right) = 1 - \alpha ,$$

wobei  $t_{r;1-\alpha/2}$  die entsprechenden Quantile der t-Verteilung mit  $r$  Freiheitsgraden bezeichnen. Je nach zugrunde gelegter Situation gelten die folgenden Standardfehler und Freiheitsgrade:

	$p$ separate Stichproben (heteroskedastisch)	aus linearem Modell (homoskedastisch)
SEM	$\hat{\sigma}_i / \sqrt{n_i}$	$\sqrt{\text{MSE}} / \sqrt{n_i}$
FG $r$	$n_i - 1$	$N - p$

Die Berechnung der Vertrauensintervalle erfolgt im linearen Modell unter Berücksichtigung der Reststreuung aller Beobachtungen, wo hingegen im heteroskedastischen Fall jedes Konfidenzintervall unabhängig von den übrigen Messreihen be-

stimmt wird. Für den Vergleich aller  $p$  Stufen mit einer „Benchmark“ (Vergleichswert, z.B. der Mittelwert aller Stufen oder auch ein anerkannter Sollwert) eignet sich also das Konfidenzintervall aus dem linearen Modell.

Die Annahme der Homoskedastizität kann mittels des Tests von Levene geprüft werden. Im Fall von Einrichtungsvergleichen wird die Durchführung dieses Tests nicht empfohlen, da eine Verletzung der Annahme einerseits schwer interpretierbar und andererseits das Ergebnis nur bei Vorliegen extremer Heteroskedastizität nennenswert beeinflusst wird.

Generell muss angemerkt werden, dass bei wachsender Anzahl von Stufen der Einflussgröße ( $p$ ) entsprechend viele solcher Konfidenzintervalle bestimmt werden müssen. Die Aussage, dass der wahre unbekannte Gruppenmittelwert  $\beta_i$  mit Sicherheit  $(1 - \alpha)$  innerhalb der Grenzen des Vertrauensintervalls liegt, gilt damit natürlich nicht mehr global, d.h. für alle  $p$  Stufen, sondern nur für jede Stufe einzeln. Um das globale Niveau  $\alpha$  einzuhalten, gibt es geeignete Adjustierungsmethoden, wie beispielsweise die von Bonferroni oder Scheffé vorgeschlagenen.

### Paarweise Gruppenvergleiche

Für zwei Gruppen lässt sich ein  $1 - \alpha$  Konfidenzintervall für den wahren Unterschied  $d$  zwischen den Mittelwerten zweier Grundgesamtheiten  $i$  und  $i'$

$$\begin{aligned} d &= \mu_i - \mu_{i'} \\ &= \beta_i - \beta_{i'} , \end{aligned}$$

der durch  $\hat{d} = \bar{y}_i - \bar{y}_{i'}$  geschätzt wird,

unbekannte aber gleiche Varianzen in beiden Gruppen (Homoskedastizität) vorausgesetzt, wie folgt angeben:

$$P \left( d \in \left[ \hat{d} - t_{n_1+n_2;1-\alpha/2} * s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} ; \hat{d} + t_{n_1+n_2;1-\alpha/2} * s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \right] \right) = 1 - \alpha .$$

Die gemeinsame Stichprobenvarianz  $s_p^2$  ist dabei gegeben als

$$s_p^2 = \frac{\sum_{i=1}^{n_1} y_{1i}^2 - \frac{1}{n_1} \left( \sum_{i=1}^{n_1} y_{1i} \right)^2 + \sum_{j=1}^{n_2} y_{2j}^2 - \frac{1}{n_2} \left( \sum_{j=1}^{n_2} y_{2j} \right)^2}{n_1 + n_2 - 2} .$$

### Betrachtungen zur Modellgüte

Zur Beurteilung, wie „gut“ die vorliegenden Daten durch das Modell beschrieben werden, dient als Kriterium das **Bestimmtheitsmaß**  $R^2$ , also das Verhältnis zwischen der durch das Modell erklärten Streuung zur Gesamtstreuung der Versuchsergebnisse  $y_{ij}$ . Im Modell der einfachen Varianzanalyse ist damit  $R^2$  gerade die (gewichtete) quadratische Streuung zwischen den vom Modell geschätzten Werten im Verhältnis zur Streuung der durch die Stichprobe gegebenen Werten:

$$R^2 = \frac{\sum_{i=1}^p \sum_{j=1}^{n_i} (\hat{y}_{ij} - \bar{y}_{..})^2}{\sum_{i=1}^p \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{..})^2}, \quad \text{mit } \hat{y}_{ij} = \hat{\alpha}_i = \bar{y}_{i.} .$$

Offensichtlich gilt:

- $R^2 = 1 - \text{SSE}/\text{SSG}$ ;
- $0 \leq R^2 \leq 1$  ;
- $R^2 = 1 \iff \text{SSE} = 0$  (Modell erklärt alles)
- $R^2 = 0 \iff \text{SSE} = \text{SSG}$  (Modell erklärt nichts).

In der einfachen Varianzanalyse gilt zusätzlich wegen  $\text{SST} + \text{SSE} = \text{SSG}$

$$R^2 = \frac{\text{SST}}{\text{SSG}} .$$

$R^2$  wird häufig auch als „Modellgüte“, „Anpassungsgüte“ oder „goodness of fit“ bezeichnet.

Der Pearsonsche (bivariate) Korrelationskoeffizient  $r_{Y, \hat{Y}}$  zwischen den Merkmalen  $Y$  und  $\hat{Y}$ , also zwischen den tatsächlich beobachteten Werten  $y_{ij}$  und den vom Modell geschätzten Werten  $\hat{y}_{ij}$ , ist gerade

$$\begin{aligned}
r_{Y,\hat{Y}} &= \frac{\sum_{i=1}^p \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{..})(\hat{y}_{ij} - \bar{y}_{..})}{\sqrt{\sum_{i=1}^p \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{..})^2 * \sum_{i=1}^p \sum_{j=1}^{n_i} (\hat{y}_{ij} - \bar{y}_{..})^2}} \\
&= \frac{\sum_{i=1}^p \sum_{j=1}^{n_i} y_{ij} \hat{y}_{ij} - N \bar{y}_{..}^2}{\sqrt{\left( \sum_{i=1}^p \sum_{j=1}^{n_i} y_{ij}^2 - N \bar{y}_{..}^2 \right) * \left( \sum_{i=1}^p \sum_{j=1}^{n_i} \hat{y}_{ij}^2 - N \bar{y}_{..}^2 \right)}} .
\end{aligned}$$

Grundsätzlich gilt für  $r_{Y,\hat{Y}}$  natürlich

$$-1 \leq r_{Y,\hat{Y}} \leq 1 ,$$

aber wegen der Schätzeigenschaften der Modellparameter gilt zusätzlich

$$0 \leq r_{Y,\hat{Y}} \leq 1 .$$

Außerdem gilt:

$$r_{Y,\hat{Y}} = \sqrt{R^2} .$$

## Diskussion

Das beschriebene Modell hat in der praktischen Anwendung den Vorteil, dass die Berechnung der Effekte einfach durchzuführen und die Ergebnisse leicht verständlich bzw. interpretierbar sind.

In der Praxis ist allerdings häufig **keine** zufällige Zuordnung der Objekte zu der Behandlung möglich, beispielsweise durch den Umstand, dass in technisch besonders gut ausgestatteten Kliniken eher Patienten mit einem **bestimmten** (zum Beispiel schweren) Erkrankungsprofil behandelt werden. Eine der Grundannahmen für die Anwendung der Einfachklassifikation ist dann nicht erfüllt.

Ein weiterer Nachteil nicht randomisierter Versuchspläne besteht darin, dass die Variabilität innerhalb der Gruppen recht groß sein kann, etwa durch therapeutisch

bedeutsame Subgruppen bedingte Heterogenität der Gesamtpopulation (d.h. durch Parameter, die Einfluss auf das Zielkriterium besitzen). Hierdurch können mögliche Unterschiede zwischen den Stufen nicht erkannt werden, da sich mehrere Effekte überlagern können. So könnte etwa eine Behandlungsart, die nur Patienten mit schwerem akutem Myokardinfarkt verordnet wird, bei einem möglichen Vergleich zwischen Behandlungen hinsichtlich der Überlebenschancen schlecht abschneiden, da Patienten mit schwerwiegenderem Infarkt eine schlechtere Prognose besitzen, unabhängig von der Behandlung. Einen ähnlichen Effekt könnte man für medizinische Einrichtungen, die beispielsweise durch unterschiedliche Ausstattung unterschiedliche Patientengruppen behandeln, vermuten.

Daher wird es bei Einrichtungsvergleichen immer erforderlich sein, von diesem einfachen („naiven“) Mittelwert-Vergleich Abstand zu nehmen und auf Modelle zurückzugreifen, die mehr als nur einen Faktor bzw. die hierarchische Struktur der Situation („Patient in Klinik“) berücksichtigen.

### 2.2.1.2 Blockexperimente

Eine Möglichkeit, die Unterschiede zwischen den Gruppen besser zu ermitteln, ist die Durchführung eines (einfachen) „Blockexperiments“, falls (eine) Hauptstörvariable(n) bekannt ist/sind. Ein „Block“ bezeichnet dabei eine Ausprägung der Störgröße.

Werden wieder  $p$  Behandlungen untersucht und liegt die Störvariable in  $r$  Ausprägungen (Blöcken) vor, wird jede Behandlung in jedem Block genau einmal angewendet.

Das Modell hat die folgende Gestalt:

$$y_{ij} = \mu + \alpha_i + \beta_j + e_{ij} \quad \text{mit} \quad i = 1, \dots, p \quad \text{und} \quad j = 1, \dots, r,$$

wobei  $y_{ij}$  das Versuchsergebnis des  $j$ -ten Blocks in der  $i$ -ten Behandlung,  $\mu$  wieder den Gesamtmittelwert und die  $\alpha_i$  die Gruppeneffekte mit

$$\sum_{i=1}^p \alpha_i = 0$$

bezeichnen. Der Term  $\beta_j$  beschreibt die  $j$ -te Ausprägung der Störvariablen, und auch hier gilt

$$\sum_{j=1}^r \beta_j = 0$$

als Reparametrisierungsbedingung.

Für die Gesamtfallzahl des Experiments gilt im balancierten Fall stets  $N = pr$ .

Die Parameter  $\mu$ ,  $\alpha_i$  und  $\beta_j$  werden geschätzt durch:

$$\begin{aligned}\hat{\mu} &= m = \bar{y}_{..} \quad , \\ \hat{\alpha}_i &= a_i = \bar{y}_{i.} - \bar{y}_{..} \text{ und} \\ \hat{\beta}_j &= b_j = \bar{y}_{.j} - \bar{y}_{..} \text{ ,}\end{aligned}$$

mit

$$\begin{aligned}\bar{y}_{i.} &= \frac{1}{r} \sum_{j=1}^r y_{ij} \quad , \\ \bar{y}_{.j} &= \frac{1}{p} \sum_{i=1}^p y_{ij} \quad \text{und} \\ \bar{y}_{..} &= \frac{1}{rp} \sum_{i=1}^p \sum_{j=1}^r y_{ij} \quad .\end{aligned}$$

Die Analysetafel für das einfache Blockexperiment hat folgende Gestalt:

Die Abkürzungen, die bereits in Tabelle 2.1 verwendet wurden, gelten auch hier. Zusätzlich bezeichnen:

SSB            Sum of Squares for Blocks (Quadratsumme der Blöcke),  
MSB            Mean Square for Blocks (Mittlere Quadratsumme der Blöcke).

Möchte man nun prüfen, ob tatsächlich signifikante Unterschiede zwischen den  $p$  Behandlungsstufen bestehen, wird wieder folgender Hypothesensatz (vgl. Abschnitt 2.2.1.1, Seite 32) zugrunde gelegt:

$$\begin{aligned}H_0^a &: \alpha_i = 0 \quad \text{für alle } i \in (1, \dots, p) \quad \text{vs.} \\ H_1^a &: \alpha_i \neq 0 \quad \text{für wenigstens ein } i \in (1, \dots, p) \text{ ,}\end{aligned}$$

**Tabelle 2.2: Analysetafel des einfachen Blockexperiments**

Streuungs- ursache	Freiheits- grade	Quadratsumme	Mittlere Quadratsumme
Unterschiede zwischen den Behandlungen	$p - 1$	$SST = \sum_{i=1}^p a_i^2$	$MST = \frac{SST}{p - 1}$
Unterschiede zwischen den Blöcken	$r - 1$	$SSB = \sum_{j=1}^r b_j^2$	$MSB = \frac{SST}{r - 1}$
zufälliger Fehler	$(p-1)(r-1)$	$SSE = SSG - SSA - SSB$	$MSE = \frac{SSE}{(p-1)(r-1)}$
Gesamt	$rp - 1$	$SSG = \sum_{i=1}^p \sum_{j=1}^r (y_{ij} - \bar{y}_{..})^2$	

mit  $\alpha_i$ : wahrer unbekannter Effekt der  $i$ -ten Behandlung.

Und entsprechend gilt für die  $r$  Blockeffekte

$$\begin{aligned}
 H_0^b: & \quad \beta_j = 0 && \text{für alle } j \in (1, \dots, r) && \text{vs.} \\
 H_1^b: & \quad \beta_j \neq 0 && \text{für wenigstens ein } j \in (1, \dots, r),
 \end{aligned}$$

mit  $\beta_j$ : wahrer unbekannter Effekt des  $j$ -ten Blocks (also der  $j$ -ten Ausprägung der Störvariablen).

Zur Prüfung von  $H_0^a$  (mit vorgegebener Irrtumswahrscheinlichkeit  $\alpha_a$ ) wird die Prüfgröße (Teststatistik)  $F$  aus den Werten MST und MSE aus obiger Tabelle gebildet. Diese kann als unter der Nullhypothese  $F$ -verteilt mit  $p - 1$  und  $(r - 1)(p - 1)$  Freiheitsgraden angenommen werden:

$$F_a = \frac{MST}{MSE} \sim F_{p-1, (r-1)(p-1)} .$$

Die kritische Grenze  $c_{\alpha_a}$  zur Definition des Ablehnungsbereichs wird entsprechend gebildet als:

$$c_{\alpha_a} = F_{p-1, (r-1)(p-1), 1-\alpha_a} .$$

Für das Testen von  $H_0^b$  (mit vorgegebener Irrtumswahrscheinlichkeit  $\alpha_b$ ) werden die Terme MSB und MSE aus obiger Tabelle benutzt.  $F_b$  kann angenommen werden als unter der Nullhypothese  $F$ -verteilt mit  $r - 1$  und  $(r - 1)(p - 1)$  Freiheitsgraden:

$$F_b = \frac{\text{MSB}}{\text{MSE}} \sim F_{r-1, (r-1)(p-1)} .$$

Für die kritische Grenze  $c_{\alpha_b}$  zur Testentscheidung gilt:

$$c_{\alpha_b} = F_{r-1, (r-1)(p-1), 1-\alpha_b} .$$

## Diskussion

Blockexperimente spielen in der medizinischen Versorgungsforschung keine bedeutende Rolle, da sich die Versuchsanordnung in der Praxis oft nicht durchführen lässt.

Im Beispiel des akuten Myokardinfarkts ist es nicht denkbar, einen Patienten mehreren (bzw. allen  $r$  verfügbaren) Ausprägungen der Störgröße zuzuweisen, da die Patienten diese Größe oft selbst tragen. Wenn die Störvariable andererseits durch eine Einrichtung (Klinik) gegeben ist, d.h. sie wird nicht vom Patienten getragen, sondern ihm/ihr zugewiesen, ist dennoch oft kein Blockexperiment durchführbar, da bei schweren Erkrankungen oft nur ein Behandlungsversuch unternommen werden kann.

In der Psychologie findet das Blockexperiment jedoch durchaus Anwendung, beispielsweise wenn man den Einfluss von Rhythmen auf das Sprachverhalten bspw. von Stotterern untersuchen will. Hier kann man die Rhythmusart als festen Behandlungseffekt und das Subjekt als Störgröße modellieren (falls die Unterschiede in der Grundgesamtheit nicht von Interesse sind, vgl. Abschnitt 2.3.1 Modelle mit zufälligen Effekten). Jede Versuchsperson muss dann zu jedem eingespielten Rhythmus Testtexte sprechen.

### 2.2.1.3 Zweifache Varianzanalyse

Besser als im Blockexperiment lassen sich Behandlungseffekte feststellen, wenn mehrere Beobachtungen pro Kombination aus Behandlung und Block vorliegen (eine solche Faktorstufenkombination nennt man auch „Zelle“). Kann man *Zelle* und

*Störgröße* eindeutig als einen Blockfaktor betrachten, d.h. wenn eine mehrfache Zellenbesetzung lediglich einen Wiederholungseffekt bewirkt, geht man von keinerlei Wechselwirkungen zwischen den beiden Faktoren aus (Beispiel: Labor  $j$  untersucht  $p$  Zigarettensorten  $n$ -mal).

Liegen die Daten jedoch so vor, dass jede Faktorstufenkombination denkbar ist (alle Zellen sind besetzt) und dass keiner der beiden Faktoren als Blockfaktor angesehen werden kann, geht man von einem Modell mit Wechselwirkungen aus.

Zunächst wird hier das Modell **ohne** Wechselwirkung dargestellt.

Werden wieder  $p$  Behandlungen untersucht und liegt die Störvariable in  $r$  Ausprägungen vor, werden in jeder Zelle  $q$  Beobachtungen bzw. bei nicht balancierten Versuchen  $n_{ij}$  Beobachtungen gemacht.

Das Modell der zweifachen Varianzanalyse (ohne Wechselwirkung) hat die folgende Gestalt:

$$y_{ijk} = \mu + \alpha_i + \beta_j + e_{ijk} \quad \text{mit} \quad i = 1, \dots, p, \quad j = 1, \dots, r \quad \text{und} \quad k = 1, \dots, q,$$

wobei  $y_{ijk}$  das Versuchsergebnis der  $k$ -ten Wiederholung innerhalb der Faktorstufenkombination (Zelle,  $j$ -ter Block in der  $i$ -ten Behandlung),  $\mu$  wieder den Gesamtmittelwert und die  $\alpha_i$  die Gruppeneffekte mit  $\sum_{i=1}^p \alpha_i = 0$  bezeichnen. Der Term  $\beta_j$  beschreibt wieder den Effekt des  $j$ -ten Blocks, und es gilt wie schon zuvor gefordert  $\sum_{j=1}^r \beta_j = 0$  als Reparametrisierungsbedingung.

Für die Gesamtfallzahl des Experiments gilt stets

$$N = prq \quad \text{bzw. bei nicht balancierten Versuchen} \quad N = \sum_{i=1}^p \sum_{j=1}^r n_{ij}.$$

Die Parameter  $\mu$ ,  $\alpha_i$  und  $\beta_j$  werden im Falle einer balancierten Situation geschätzt durch:

$$\begin{aligned}\hat{\mu} &= m = \bar{y}_{...} && , \\ \hat{\alpha}_i &= a_i = \bar{y}_{i..} - \bar{y}_{...} && \text{und} \\ \hat{\beta}_j &= b_j = \bar{y}_{.j.} - \bar{y}_{...} && ,\end{aligned}$$

mit

$$\begin{aligned}\bar{y}_{i..} &= \frac{1}{rq} \sum_{j=1}^r \sum_{k=1}^q y_{ijk} && , \\ \bar{y}_{.j.} &= \frac{1}{pq} \sum_{i=1}^p \sum_{k=1}^q y_{ijk} && \text{und} \\ \bar{y}_{...} &= \frac{1}{prq} \sum_{j=1}^r \sum_{i=1}^p \sum_{k=1}^q y_{ijk} && .\end{aligned}$$

Auf die Darstellung der Analysetafel, der Punktschätzungen und der Hypothesentests wird aufgrund der Analogie zu den vorherigen Abschnitten verzichtet.

## Diskussion

Häufig kommen Versuchspläne der zweifachen Varianzanalyse (meist mit Wechselwirkungen) in klinischen Studien zum Einsatz. Ist beispielsweise die Störvariable vor Versuchsbeginn bekannt, kann diese als Stratifizierungsfaktor (auch „Schichtungsfaktor“ genannt) benutzt werden, wodurch innerhalb eines jeden Stratums (jeder Schicht) die Zuweisung zur Behandlung zufällig erfolgt. In der Regel muss vor Studienbeginn davon ausgegangen werden, dass möglicherweise Wechselwirkungen zwischen der Stratifizierungsvariablen und der Behandlung bestehen, d.h. dass sich die Wirksamkeit einer Testsubstanz im Vergleich zu einer Kontrollbehandlung zwischen den Untergruppen unterscheidet.

### 2.2.2 Regressionsanalyse

Bei der Regressionsanalyse liegen die Faktoren nicht als gruppierte, sondern als kontinuierliche („stetige“) Variablen vor. Für die Elemente der Designmatrix  $X$  gilt dann

$$x_{ij} \in \{0, 1\} \quad \text{für genau ein } i \quad \text{mit} \quad j = 1, \dots, k .$$

Das heißt, höchstens ein Vektor enthält ausschließlich Elemente mit dem Wert 0 oder 1.

Im Fall, dass der Einfluss genau **eines** stetigen Faktors auf die Zielgröße untersucht werden soll, spricht man von „einfacher Regressionsanalyse“; Modelle in denen **mehrere** stetige Einflussgrößen modelliert werden, heißen allgemein „multiple Regressionsanalyse“.

### Beispiel für eine einfache Regressionsanalyse

Es soll der Einfluss der Dosis eines Arzneimittels auf den arteriellen Blutdruck untersucht werden. Der Blutdruck von  $n$  untersuchten Patienten ist dann im Vektor  $y = (y_i)$  (mit  $i = 1, \dots, n$ ) realisiert. Der Vektor  $\beta = (\beta_0, \beta_1)$  der Einflussfaktoren enthält ein Absolutglied  $\beta_0$  und den Steigungskoeffizienten  $\beta_1$ .

Das Modell lässt sich dann in Matrixschreibweise darstellen als

$$y = X\beta + e = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix} = \begin{pmatrix} 1 & X_{11} \\ 1 & X_{12} \\ \vdots & \vdots \\ 1 & X_{1N} \end{pmatrix} \begin{pmatrix} \beta_0 & \beta_1 \end{pmatrix} + \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_N \end{pmatrix}$$

bzw. als klassische Regressionsgleichung in Summationsschreibweise

$$y_i = \beta_0 + \beta_1 x_{1i} + e_i \quad \text{mit} \quad i = 1, \dots, N.$$

### 2.2.3 Kovarianzanalyse

Die Kovarianzanalyse kann als Kombination aus Varianz- und Regressionsanalyse verstanden werden. Das heißt, die Einflussfaktoren liegen sowohl in gruppierter als auch in stetiger Form vor. Mindestens ein Vektor der Matrix  $X$  enthält ausschließlich Elemente mit Werten von 0 oder 1, jedoch nicht alle.

Ein einfaches Kovarianzanalyse-Modell schreibt sich dann in der Summationsdarstellung als

$$y_{ij} = \mu + \alpha_i + \beta x_{ij} + e_{ij} \quad \text{mit} \quad i = 1, \dots, k \quad \text{und} \quad j = 1, \dots, n_i.$$

## 2.2.4 Generalisierte Lineare Modelle – Logistische Regressionsanalyse

Bisher wurden Modelle vorgestellt, bei denen die abhängige Variable  $Y$  stetiges Skalenniveau besitzt. Ein Spezialfall der linearen Regression ist die „logistische Regression“, bei der eine *binäre* Variable  $Y$  als Zielgröße betrachtet wird. Bei binären (auch „dichotom“ genannten) Daten interessiert man sich meist für das Auftreten eines bestimmten Ereignisses und dessen Wahrscheinlichkeit, welches sowohl positiv (z.B. durch erfolgte „Heilung“) als auch negativ (z.B. durch „Tod“) besetzt sein kann.

Die Besetzung der beiden möglichen Ausprägungen der Zielgröße werden im Folgenden stets wie nachstehend bezeichnet:

- $Y = 1$ : „das interessierende Ereignis tritt ein“, oder
- $Y = 0$ : „das interessierende Ereignis tritt nicht ein“.

Bei der Modellierung wird also stets die Wahrscheinlichkeit für das Auftreten des interessierenden Ereignisses und nicht dessen Komplementäreignis betrachtet. Im Fall des Behandlungserfolgs (ja/nein) interessiert man sich typischerweise für die Wahrscheinlichkeit eines *Erfolgs*; also bezeichnet  $P(Y = 1)$  die Erfolgswahrscheinlichkeit. Bei negativer Konnotation des interessierenden Ereignisses würde  $Y = 1$  entsprechend das negative Ereignis bezeichnen.

Häufig werden Daten, die ursprünglich in einem höheren Skalenniveau vorliegen, durch eine fest definierte Vorschrift nachträglich „dichotomisiert“, wie beispielsweise bei einer Blutdrucksenkung:

$$Y_i = \begin{cases} 1 \text{ (Erfolg)} & \text{falls Blutdruck des } i\text{-ten Patienten } z_i \text{ bei Entlassung } \leq c, \\ 0 \text{ (Misserfolg)} & \text{falls Blutdruck des } i\text{-ten Patienten } z_i \text{ bei Entlassung } > c. \end{cases}$$

Obwohl bei einer solchen Transformation Information verloren geht, wird eine Dichotomisierung häufig der Betrachtung der Ursprungsdaten vorgezogen, beispielsweise weil in der Öffentlichkeit Raten (%-Angaben) besser verstanden werden als der Vergleich von Mittelwerten.

Das Modell der logistischen Regression wird in der folgenden Form dargestellt:

$$P(Y = 1) = \frac{1}{1 + e^{-X}} \quad \text{mit} \quad X = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k .$$

Der Zusammenhang zwischen den Einflussgrößen und der Zielgröße ist somit zumindest in  $X$  linear. Möchte man die Faktoren des Modells als Linearkombination wie im konventionellen linearen Modell als  $X\beta + e$  darstellen, kann dies über eine Funktion  $g$  geschehen.

Die Funktion

$$\text{logit}(p) = \ln \left( \frac{p}{1-p} \right) = X\beta + e$$

( $p$  bezeichnet hier der Einfachheit halber  $P(Y = 1)$ ) stellt einen solchen Zusammenhang her. Ein logistisches Regressionsmodell, so dargestellt, wird als „Logit-Modell“ bezeichnet; das Verhältnis zwischen Wahrscheinlichkeit  $p$  und Gegenwahrscheinlichkeit  $(1-p)$  heißt „Odds“.  $\text{Logit}(p)$ , also der natürliche Logarithmus des Odds, nennt man das „Log Odds“.

Das Logit-Modell lässt sich als ein GENERALISIERTES LINEARES MODELL (GLM) auffassen, nämlich in der Weise, dass eine Funktion  $g = g(p)$  der Zielgröße als *lineare* Kombination der Einflussgrößen in  $X$  modelliert wird. Da  $g$  eine Beziehung zwischen den Einflussgrößen  $X$  und der stochastischen Größe  $Y$  herstellt, wird diese häufig als „Link-Funktion“ bezeichnet (siehe Nelder und Wedderburn [41]). Das GLM schreibt sich dann allgemein zu

$$g(p) = X\beta + e .$$

Andere Link-Funktionen wie die „Probit“-Funktion  $g(p) = \Phi^{-1}(p)$  oder andere stehen für binäre Zielgrößen zur Verfügung.

Das logistische Regressionsmodell lässt sich auch für kategorielle Einflussgrößen (wie beispielsweise Geschlecht, Behandlung oder auch Klinikum) verwenden. Liegt eine Variable  $X_i$  in  $p$  Stufen ohne interne Ordnung vor, so lassen sich ihre Stufen

durch Bildung von  $k$  Designvariablen  $X_{i1}, \dots, X_{ip}$  mit jeweils  $p$  Einträgen, also in einer  $k \times k$ -Matrix, beispielsweise der folgenden oder einer ähnlichen Form

Stufe	$X_{i1}$	$X_{i2}$	$X_{i3}$	$\dots$	$X_{ip}$
1	1	0	0	$\dots$	0
2	0	1	0	$\dots$	0
3	0	0	1	$\dots$	0
$\vdots$			$\vdots$		
$p$	0	0	0	$\dots$	1

parametrisieren. Der  $j$ -ten Stufe der Variablen  $X_i$  wird somit in Analogie zur Varianzanalyse eine solche Designvariable  $X_{ij}$  mit jeweils einem Koeffizientenschätzer  $\beta_{ij}$  zugewiesen. Im Beispiel der einfachen logistischen Regression schreibt sich das Logit von  $p$  auf der  $j$ -ten Stufe ( $g(p_{ij}) = \text{logit}(p_{ij})$ ) dann mit der oben gewählten Parametrisierung zu

$$g(p_{ij}) = \beta_0 + \sum_{l=1}^p \beta_{il} \mathbb{I}(X_{il} = 1) = \beta_0 + \beta_{ij} .$$

$p_j$  bezeichnet wieder die Wahrscheinlichkeit des interessierenden Ereignisses, falls  $X_i = j$ , also  $P(Y = 1 | X_i = j)$ .

### Anteils- und Parameterschätzung

Ähnlich wie bei stetigen Zielgrößen (Varianzanalyse, Regressionsanalyse) möchte man auch bei binären Zielgrößen Schätzer für die Parameter im Modell angeben. Falls eine Einflussgröße  $X_i$  stetig ist, erhält man einen linearen Zusammenhang zwischen  $g(p)$  und dem Regressor  $\beta_i X_i$ . Falls die Einflussgröße in Stufen – und nicht als stetige Zielgröße – vorliegt, sollen Schätzer für die Wahrscheinlichkeiten des interessierenden Ereignisses auf jeder Stufe  $j$  über  $\beta_{ij} X_{ij}$  angegeben werden. Wird in einer Modellanpassung nur ein Einflussparameter  $X_1$  betrachtet und ist dieser – wie bei Einrichtungsvergleichen denkbar – eine kategorielle Größe, kann man von einer einfachen logistischen ANOVA sprechen, deren Daten sich mittels  $2 \times p$ -Kontingenztafeln darstellen lassen.

Wird die obige Modellgleichung mit  $X\beta = \beta_0 + \beta_1 X_1$  angenommen, so können die relativen Häufigkeiten des interessierenden Ereignisses ( $Y = 1$ ) auf den  $k$  Stufen des Faktors  $X_1$

$$h_i = \frac{H_i}{n_i} = \frac{\sum_{j=1}^{n_i} y_{ij}}{n_i} = \hat{p}_i$$

$$\text{mit } y_{ij} = \begin{cases} 1 & \text{falls das Ereignis bei Subjekt } j \text{ in Gruppe } i \text{ eintritt} \\ 0 & \text{falls das Ereignis bei Subjekt } j \text{ in Gruppe } i \text{ nicht eintritt} \end{cases}$$

als erwartungstreue Schätzer für die wahren Ereigniswahrscheinlichkeiten  $p_i$ , ( $i = 1, \dots, k$ ) bestimmt werden. Diese lassen sich mittels der gewählten Link-Funktion  $g(p)$  und der gewählten Parametrisierung von  $X_1$  in die entsprechenden Koeffizienten in  $X$  überführen.

Im einfachsten möglichen Beispiel ( $p = 2$ ) ergibt sich eine  $2 \times 2$ -Kontingenztafel. Da  $X_1$  und  $X_2$  bekannt sind, existiert jedoch aufgrund der Beziehungen

$$\begin{aligned} X_1 &= \beta_0 + \beta_{11} \\ X_2 &= \beta_0 + \beta_{12} \end{aligned}$$

keine eindeutige Lösung für die Koeffizienten  $\beta_0$ ,  $\beta_{11}$  und  $\beta_{12}$ . Da aber zumindest  $\beta_{11} - \beta_{12} = X_1 - X_2$  bekannt ist, kann man durch das Festhalten von  $\beta_{12} \equiv 0$  eine Lösung bestimmen:

$$\beta_0 = X_2 \quad \text{und} \quad \beta_1 = X_1 - X_2 .$$

Werden die beobachteten Häufigkeiten wie oben bezeichnet als Anteilsschätzer benutzt – d.h. die  $p$  Stichproben werden separat betrachtet – gilt für die Varianz der beobachteten Anteilsschätzer  $\hat{p}_i$ :

$$\text{Var}(\hat{p}_i) = \frac{p_i(1-p_i)}{N} \quad \text{bzw.} \quad \widehat{\text{Var}}(\hat{p}_i) = \frac{\hat{p}_i(1-\hat{p}_i)}{N} .$$

Konfidenzintervalle für die  $p_i$  können unter Verwendung des Standardfehlers der Anteilsschätzungen

$$\hat{\sigma}(\hat{p}_i) = \sqrt{\widehat{\text{Var}}(\hat{p}_i)}$$

mit Abdeckungssicherheit  $(1 - \alpha)$  zu

$$P \left( p_i \in \left[ \hat{p}_i - u_{\alpha/2} \hat{\sigma}(\hat{p}_i) ; \hat{p}_i + u_{\alpha/2} \hat{\sigma}(\hat{p}_i) \right] \right) = 1 - \alpha$$

angegeben werden, wobei  $u_\alpha$  die  $\alpha$ -Quantile der Standardnormalverteilung bezeichnen.

Allgemein werden im logistischen Modell die Schätzungen für die Koeffizienten  $\hat{\beta}$  zur Schätzung der Wahrscheinlichkeiten für alle Kombinationen in  $X$  verwendet. Konfidenzintervalle für die Parameterschätzer  $\hat{\beta}_i$  können z.B. nach Wald, basierend auf dem Standardfehler  $\hat{\sigma}(\hat{\beta}_i)$ , der sich als die Wurzel des  $i$ -ten Diagonalelements der geschätzten Varianz-Kovarianzmatrix  $\hat{V}$  der Parameter und der asymptotischen Normalverteilung der Schätzungen berechnet, bestimmt werden zu

$$\hat{\beta}_i \pm u_{1-\alpha/2} \hat{\sigma}(\hat{\beta}_i) .$$

Alternativ hierzu gibt es „Profile Likelihood“-Konfidenzintervalle auf Basis des verallgemeinerten Likelihood-Ratio-Tests (siehe auch Venzon and Moolgavkar [56]).

Für die mittels der Koeffizienten-Schätzer bestimmten Ereigniswahrscheinlichkeiten

$$\hat{p}_i(Y = 1) = \frac{1}{1 + e^{-\hat{X}}} = \frac{1}{1 + e^{-(\hat{\beta}_0 + \hat{\beta}_1 X_1)}}$$

können nun ebenfalls unter Bestimmung des geschätzten Standardfehlers von  $\hat{\beta}$  durch

$$\hat{\sigma}(\hat{X}) = \sqrt{(1, X') \hat{V} (1, X)'}$$

( $\hat{V}$  ist die geschätzte Kovarianzmatrix der Parameterschätzer), angegeben werden zu

$$P \left( p_i(Y = 1) \in \left[ \frac{1}{1 + e^{-\hat{X} - u_{\alpha/2} \hat{\sigma}(\hat{X})}} ; \frac{1}{1 + e^{-\hat{X} + u_{\alpha/2} \hat{\sigma}(\hat{X})}} \right] \right) = 1 - \alpha .$$

Die so erzeugten Konfidenzintervalle sind in aller Regel - insbesondere bei balancierten Versuchen - deutlich kleiner als die separat pro Stufe bestimmten Bereiche.

### Inferenz in der logistischen Regression

Bei der logistischen Regression interessieren beim Vergleich zwischen zwei Stufen  $j$  und  $k$  ( $j \neq k$ ) einer Einflussgröße  $X_i$  mehrere Kenngrößen:

- Risiko-Differenz (RD)  $p_{ij} - p_{ik}$ ,
- Relatives Risiko (RR)  $p_{ij}/p_{ik}$ ,
- Odds Ratio (OR)  $\Psi_{i_{jk}} = O_{ij}/O_{ik}$ , mit  $O_{il} = p_{il}/(1 - p_{il})$ ,
- Log Odds Ratio (LOR)  $\theta_{i_{jk}} = \ln(\Psi_{i_{jk}}) = \ln(O_{ij}) - \ln(O_{ik})$ .

Unter Verwendung der Anteilsschätzer  $\hat{p}_i$  bzw.  $\hat{p}_j$  ergeben sich Schätzwerte für die oben genannten Kriterien. Unter der Nullhypothese  $H_0 : p_i = p_j$  haben RR und OR den Erwartungswert 1, RD und LOR den Erwartungswert 0. Im Beispiel der einfachen logistischen Regression gilt also für LOR zwei beliebiger Ausprägungen einer Einflussgröße  $X_i$ ,  $X_{ij}$  und  $X_{ik}$  in der Schreibweise der Logit-Funktion  $g(p) = \ln(\text{logit}(p))$ :

$$\theta_{i_{jk}} = \ln(\Psi_{i_{jk}}) = g(p_{ij}) - g(p_{ik}) .$$

Im Rahmen dieser Arbeit wird in erster Linie das Odds Ratio bzw. das Log Odds Ratio (LOR) als Kriterium betrachtet werden, da sie sich im Logit-Modell aus den Schätzern der Modellkoeffizienten direkt berechnen lassen.

Allgemein können auf Basis der (Maximum-Likelihood-)Schätzungen (eine Einführung von Maximum-Likelihood-Schätzungen wird in Kapitel 2.3.1 gegeben) für  $\hat{\beta}_i$  im generalisierten linearen Modell beliebige geschätzte Log Odds Ratios für ein Paar ( $X_i = a, X_i = b$ ) einfach bestimmt werden zu:

$$\begin{aligned} \hat{\Psi}_{i_{ab}} &= e^{\hat{\beta}_i(a-b)} && \text{bzw.} \\ \hat{\theta}_{i_{ab}} &= \hat{g}(p_{ia}) - \hat{g}(p_{ib}) = \hat{\beta}_i(a - b) . \end{aligned}$$

Das Odds Ratio für  $a - b = 1$  kann somit bei stetigen  $X_i$  interpretiert werden als die Änderung des Odds für jede Zunahme von  $X_i$  um eine Einheit.

Bei gruppierten Einflussgrößen (ohne innere Ordnung) sind diese Kenngrößen in Abhängigkeit der Parametrisierung (siehe Seite 48) definiert. Wie weiter oben beschrieben, wird häufig der Koeffizient für die letzte Ausprägung  $\hat{\beta}_{ip}$  gleich *null* angenommen. Hier gelten die (Log) Odds Ratios analog zum steigen Fall mit  $a - b = 1$ .

Konfidenzintervalle für die Schätzer  $\hat{\Psi}$  und  $\hat{\theta}$  können aus den Parameterschätzern  $\hat{\beta}_i$  und deren geschätzter Varianz-Kovarianzmatrix  $\hat{V}$  (Fisher Information) abgeleitet werden, da jedes (Log) Odds Ratio, wie oben dargestellt, lineare Funktionen aus den Parameterschätzern selbst sind. Konfidenzintervalle nach Wald (unter Normalverteilungs-Annahme von  $\theta$ ) oder auf Basis der Likelihood Ratio Tests stehen zur Verfügung.

Zur Prüfung der Nullhypothese

$$H_0 : \Psi = 1 \text{ bzw. } \theta = 0 ,$$

d.h. um zu prüfen, ob zwischen den interessierenden Stufen tatsächlich Unterschiede in  $g(p)$  bestehen, existieren entsprechend der Konfidenzintervalle Wald-Tests bzw. Likelihood-Ratio Tests sowie Score-Tests. Für eine beliebige Linearkombination  $L(\beta)$ , also (Log) Odds Ratios zwischen beliebigen Stufen, wird die Nullhypothese allgemein geschrieben zu

$$H_0 : L\beta = c ,$$

wobei  $c$  eine beliebige Konstante darstellt. Die Wald-Statistik

$$T_W = (L\hat{\beta} - c)' \left( L(\hat{V})L' \right)^{-1} (L\hat{\beta} - c)$$

folgt unter  $H_0$  einer  $\chi^2$ -Verteilung mit  $Rg(L)$  Freiheitsgraden.

Bei kleineren Fallzahlen bzw. nicht gesicherter Normalverteilungsannahme stehen alternativ Score Tests und Likelihood-Ratio Tests zur Verfügung.

### Modellgüte in der logistischen Regression

Analog zu klassischen Varianzanalyse-Modellen ( $R^2$ ) kann im logistischen Modell ein ähnlicher Ansatz zur Bestimmung der „Goodness of Fit“ benutzt werden. Cox und Snell [7] schlugen den folgenden Koeffizienten für das verallgemeinerte lineare Modell vor:

$$R^2 = 1 - \sqrt[n]{\left(\frac{L(0)}{L(\hat{\beta})}\right)^2}.$$

$L(0)$  bezeichnet hier die Likelihood-Funktion mit lediglich dem Absolutglied (Intercept)  $\beta_0$ ;  $L(\hat{\beta})$  bezeichnet die Likelihood-Funktion des geschätzten Modells und  $n$  die Gesamtfallzahl.  $R^2$  liegt wie beim multiplen Korrelationskoeffizienten stets zwischen 0 und 1 und nimmt im Falle, dass durch die Parameter in  $\beta$  kein Beitrag zur Prediktion der Ereigniswahrscheinlichkeiten geleistet wird, einen Wert nahe 0 an.

Darüber hinaus wird bei der logistischen Regression das Konzept der konkordanten Paare zur Modellgüte betrachtet. Hier werden die vom Modell geschätzten Ereigniswahrscheinlichkeiten  $\hat{p}_i$  sämtlicher Paare von Beobachtungen mit unterschiedlicher Ausprägung der Zielgröße miteinander verglichen. Bei binärer Zielgröße existieren demnach stets

$$n(c_{ij}) = \sum y_{ij} * \left(N - \sum y_{ij}\right)$$

Paare.

Ein solches Paar  $c_{ij}$  mit  $i \neq j$   $y_i > y_j$ , also im binären Fall  $y_i = 1$  und  $y_j = 0$ , wird dann wie folgt klassifiziert:

$$c_{ij} = \begin{cases} 1 & \text{falls } \hat{p}_i > \hat{p}_j \text{ („konkordant“)} \\ 0 & \text{falls } \hat{p}_i = \hat{p}_j \text{ („unbestimmt“)} \\ -1 & \text{falls } \hat{p}_i < \hat{p}_j \text{ („diskordant“)} \end{cases}$$

Seien nun

- $N$ : Anzahl der Subjekte (Gesamtfallzahl);
- $n(c_{ij})$ : Anzahl der Beobachtungspaare mit unterschiedlicher Response;
- $n_k$ : Anzahl der konkordanten Paare;
- $n_u$ : Anzahl der unbestimmten Paare;
- $n_d$ : Anzahl der diskordanten Paare.

Zur Bestimmung der Vorhersagekraft des Modells werden die folgenden Kenngrößen diskutiert:

$$\begin{aligned}
 c: & & C &= (n_k + 0,5(n(c_{ij}) - n_c - n_d)) / (n(c_{ij})) , \\
 \text{Sommer's D:} & & D &= (n_k - n_d) / (N(c_{ij})) , \\
 \text{Goodman-Kruskal Gamma:} & & \gamma &= (n_k - n_d) / (n_k + n_d) , \\
 \text{Kendall's Tau - a} & & \tau &= (n_k - n_d) / ((N - 1)N/2) ,
 \end{aligned}$$

Die Kenngröße  $C$  gibt bei binärer Zielgröße eine Schätzung über die Fläche unter der Operationscharakteristik an (siehe Hanley and McNeil [19]). Eine detaillierte Diskussion der vorgestellten Modellgüte-Kriterien findet sich bei Agresti [1].

### Proportional Odds Models

Das Modell der logistischen Regression ist auf ordinal skalierte Parameter als Zielgröße erweiterbar. Interessiert man sich beispielsweise für den Schweregrad einer Erkrankung (wie etwa den akuten Myokardinfarkt) als Zielgröße, der in den Ausprägungen „leicht“, „mittel“ und „schwer“ gemessen wird, kann die Modellerweiterung des PROPORTIONAL ODDS MODEL benutzt werden (siehe auch Cox und Snell [7]).

Liegt die Zielgröße in  $r$  geordneten Stufen vor, wird die verallgemeinerte Modellgleichung des logistischen Modells dargestellt als  $(r - 1)$  Dichotomisierungen der Form

$$P(Y \leq i) = \frac{1}{1 + e^{-X_i}} \quad \text{mit} \quad X_i = \beta_{i0} + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_k X_{ik}$$

bzw. in der Schreibweise der Logit-Linkfunktion

$$\text{logit}(p_i) = \ln \left( \frac{p_i}{1 - p_i} \right) = X_i \quad \text{mit} \quad p_i = P(Y \leq i) \quad \text{und} \quad 1 - p_i = P(Y > i) .$$

Die Modellkoeffizienten  $(\beta_1, \dots, \beta_k)$  werden für alle  $(r - 1)$  Dichotomisierungen festgehalten, während das Absolutglied  $\beta_{i0}$  variiert. Diese Annahme der parallelen Regressionsgeraden bzw. -hyperebenen wird „Proportional Odds Assumption“ genannt und kann mittels Score-Tests überprüft werden. Falls die Annahme nicht gerechtfertigt ist, kann auf die Modelle mit diskreter Response (siehe nachfolgender Abschnitt) ausgewichen werden.

### Multinomiale Modelle

Ein weiterer Spezialfall des logistischen Modells ist gegeben, falls die Zielgröße in  $r$  Stufen ohne innere Ordnung, also auf nominalem Skalenniveau, gemessen wird. In diesem Fall – man spricht hier von MULTINOMIALEN MODELLEN oder „Discrete Choice Models“ (siehe Cox und Snell [7]) – wird die Logit-Modellgleichung verallgemeinert zu

$$\text{logit}(p_i) = \ln \left( \frac{p_i}{p_r} \right) = X_i \quad \text{mit} \quad X_i = \beta_{i0} + \beta_{1i}X_{i1} + \beta_{2i}X_{i2} + \dots + \beta_{ki}X_{ik} \quad (i=1, \dots, r-1),$$

es werden also für jede Ausprägung von  $Y$  (Log) Odds Ratios bezüglich der  $r$ -ten Ausprägung betrachtet. Die Linearisierungsvorschrift in diesem Modell wird auch „generalisiertes Logit“ („generalized logit“) genannt.

Wie zu Beginn dieses Kapitels beschrieben, sind diese Situationen im Rahmen von Einrichtungsvergleichen recht selten anzutreffen und werden daher im Rahmen dieser Arbeit nur am Rande behandelt.

#### 2.2.5 Poisson-Modelle

Bei Poisson-Modellen, die auch „Modelle der seltenen Ereignisse“ oder auch „Zählprozesse“ genannt werden, wird ein so genanntes „Urnenmodell mit Zurücklegen“ zugrunde gelegt. Das bedeutet bei Situationen mit endlichen Fallzahlen, dass ein Subjekt (Patient) das interessierende Ereignis theoretisch beliebig häufig erfahren kann.

Diese Modelle sind zur Verwendung bei Einrichtungsvergleichen ungeeignet, da Patienten das Ereignis nur einmal bzw. in begrenzter Anzahl erfahren können und nicht durch andere Patienten substituiert werden.

## 2.3 Gemischte Lineare Modelle

### 2.3.1 Einführung

Modelle, bei denen sowohl feste als auch zufällige Effekte betrachtet werden, nennt man GEMISCHTE MODELLE. Die nachfolgende Darstellung soll zur generellen Einführung der Bedeutung und der Notation bei gemischten Modellen dienen. Die Theorie der gemischten Modelle („Mixed Models Theory“) geht zurück auf die nachstehend genannten Autoren. Grundlagen der Mixed-Models Theorie sind beispielsweise bei Searle et al. [50] ausführlich beschrieben.

Das gemischte Modell schreibt sich in Anlehnung an das allgemeine lineare Modell zu

$$y = X\beta + Z\gamma$$

mit einer oder mehreren zufälligen Komponenten in  $\gamma$ .

$$E \begin{pmatrix} \gamma \\ e \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

$$D = COV \begin{pmatrix} \gamma \\ e \end{pmatrix} = \begin{pmatrix} G & 0 \\ 0 & R \end{pmatrix} .$$

$X$  bezeichnet wieder die Designmatrix für die festen Effekte  $\beta$ , und  $Z$  bezeichnet die Designmatrix der zufälligen Effekte  $\gamma$ . Da die block-diagonale Matrix  $D$  – wie hier dargestellt – lediglich aus Diagonalelementen besteht, werden also die Faktoren mit zufälligen Effekten stets unabhängig vom Restfehler  $e$  modelliert; jedoch können innerhalb der beiden Gruppen theoretisch beliebige Kovarianzstrukturen angenommen werden. Die Varianz-Kovarianzmatrix der Beobachtungen  $y = y_{ij} = (y_{11}, \dots, y_{p_{n_p}})'$  ergibt sich somit aus den Elementen von  $G$  und  $R$  zu

$$Var(y_{ij}) = V = ZGZ' + R . \quad (2.1)$$

Der Vektor der Residuen ergibt sich aus

$$r = y - X(X'V^{-1}X)^{-1}X'V^{-1}y . \quad (2.2)$$

Während für den Parameterraum der festen Effekte  $\beta$  im gemischten Modell  $\beta \in \mathbb{R}^k$  gilt, können die Parameter in  $\gamma$  nur solche Werte annehmen, für die  $G$  und  $R$  positiv (semi)definit sind.

### Schätzverfahren in gemischten Modellen

Im Vergleich zum Modell mit ausschließlich festen Effekten ist die Schätzung im gemischten Modell aufwändiger, da hier eine größere Anzahl unbekannter Parameter (simultan) zu schätzen ist, nämlich nicht nur  $\sigma_e^2$  bzw. die Elemente in  $R$ , sondern auch die Elemente in  $G$ . Kleinste-Quadrate-Methoden sind hier nicht mehr adäquat (vgl. z.B. Laird and Ware [33]).

Zur Schätzung der Variationskomponenten in  $G$  und  $R$  wird unter Benutzung der Normalverteilung-Annahme von  $\gamma$  und  $e$  zumeist mittels **Likelihood**-basierten Schätzverfahren gearbeitet. Diese Methoden basieren grundsätzlich auf dem Ansatz, diejenigen Werte für die zu schätzenden Parameter zu finden, bei denen die Wahrscheinlichkeit („likelihood“), dass eben diese Parameter-Werte als Realisationen beobachtet werden, maximal wird.

Die allgemeine Form einer Likelihood-Funktion für einen Parameter(vektor)  $\theta$  lautet bei gegebenem Beobachtungsvektor  $y$  und Dichtefunktion  $f_y$ :

$$L(\theta) = \prod_{i,j=1}^N f_y(y_{ij} | \theta) . \quad (2.3)$$

Bei der Varianz- und bei der Parameterschätzung in gemischten Modellen haben sich heute im Wesentlichen zwei Methoden etabliert:

- *Maximum-Likelihood (ML)*- und
- *Restricted Maximum-Likelihood (REML)*-Schätzung.

Allgemein gesprochen unterscheiden sich die beiden Methoden darin, dass bei der ML-Methode *alle* Parameter im Modell simultan geschätzt werden. Bei der REML-Methode hingegen werden nur die zufälligen Effekte separat, d.h. ohne Berücksichtigung der festen Effekte, geschätzt. Die festen Effekte werden dann in einem zweiten Schritt aus den ML-Schätzern bestimmt.

Patterson und Thompson [43] haben gezeigt, dass unter der Nichtbeachtung von  $\beta$  keine Information über  $\gamma$  verloren geht. Daher erfüllen beide Schätzmethode die wünschenswerten Eigenschaften der Konsistenz, Effizienz und der asymptotischen Normalität. Der Vorteil der REML-Methode liegt jedoch darin, dass im Gegensatz zur nicht restringierten ML-Methode die Schätzer der Varianzkomponenten (aufgrund der benutzten Freiheitsgrade) nicht negativ verzerrt sind. Andererseits liefern die ML-Schätzungen der Restvarianz  $\sigma_e^2$  für  $rg(X) < 5$  einen geringeren mittleren quadratischen Fehler (MSE) als die REML-Schätzer.

Zur Modellschätzung existiert darüber hinaus eine nicht-iterative Schätzmethode „MIVQUE0“ bzw. „MINQUE“ (Minimum Variance Quadratic Unbiased Estimation), die von Rao et al. [45] bzw. von Hartley et al. [20] vorgeschlagen wurde. Diese Methode liefert erwartungstreue Schätzer der Varianzkomponenten, die jedoch von den Schätzern der festen Effekte unabhängig sind. Wie von Swallow et al. [53] mittels Simulationen gezeigt wurde, besitzt die MINQUE-Methode zwar etwas schlechtere Eigenschaften als die Likelihood-basierten Methoden; auf sie kann aber im Falle von Konvergenz-Problemen bei ML und REML zurückgegriffen werden.

### Varianzschätzung im Gemischten Linearen Modell

Die Schätzung der Varianzkomponenten im Modell erfolgt in einem ersten Schritt über die Maximierung einer Log-Likelihood-Funktion für  $G$  und  $R$

$$l_{\text{ML}}(G, R) = -\frac{1}{2} \ln|V| - \frac{1}{2} r' V^{-1} r - \frac{n}{2} \ln(2\pi) ,$$

$$l_{\text{REML}}(G, R) = -\frac{1}{2} \ln|V| - \frac{1}{2} |X' V^{-1} X| - \frac{1}{2} r' V^{-1} r - \frac{N-p}{2} \ln(2\pi) .$$

Die Lösungen  $\hat{G}$  und  $\hat{R}$ , für die die Funktionen  $l_{\text{ML}}$  bzw.  $l_{\text{REML}}$  maximal werden, heißen MAXIMUM-LIKELIHOOD-SCHÄTZUNG, wobei  $p$  den Rang von  $X$ , also die Zahl der festen Faktor-Stufen-Kombinationen, bezeichnet.

Nur in sehr einfachen Modellen (d.h. Modellen mit nur einem festen Effekt und nur dem Restfehler als zufälligem Effekt) können für die Optimierungen von  $l_{\text{ML}}$  bzw.  $l_{\text{REML}}$  geschlossene Lösungen bestimmt werden; in allen anderen Fällen müssen numerische Algorithmen verwendet werden, die aber zur heutigen Zeit von den gängigen Rechnern leicht bewältigt werden können. Weite Verbreitung findet heute der (stabilisierte) Newton-Raphson-Algorithmus (siehe z.B. bei Littell et al. [35]).

### Schätzung der Effekte im Gemischten Linearen Modell

Zur Parameterschätzung im gemischten Modell wird das Gleichungssystem

$$\begin{pmatrix} X'\hat{R}^{-1}X & X'\hat{R}^{-1}Z \\ Z'\hat{R}^{-1}X & Z'\hat{R}^{-1}Z + \hat{G}^{-1} \end{pmatrix} \begin{pmatrix} \hat{\beta} \\ \hat{\gamma} \end{pmatrix} = \begin{pmatrix} X'\hat{R}^{-1}y \\ Z'\hat{R}^{-1}y \end{pmatrix}$$

unter Verwendung der wie oben bestimmten Schätzer für die Variationskomponenten nach  $\hat{\beta}$  und  $\hat{\gamma}$  aufgelöst (vgl. Henderson [24]). Die Lösungen können wie folgt dargestellt werden (vgl. Laird und Ware Henderson [33]):

$$\begin{aligned} \hat{\beta} &= (X'\hat{V}^{-1}X)^{-} X'\hat{V}^{-1}y, \\ \hat{\gamma} &= \hat{G}Z'\hat{V}^{-1}(y - X\hat{\beta}). \end{aligned}$$

$A^{-}$  bezeichnet die generalisierte Inverse einer Matrix  $A$ . Wenn  $\gamma$  als feste Effekte geschätzt wären, so stellt  $G^{-1}$  eine Schrumpfung („shrinkage“) der Schätzer in  $\gamma$  von beobachteten Mittelwerten hin zum geschätzten Populationsmittelwert  $\hat{\mu}$  dar.

Wenn die Variationskomponenten in  $G$  bekannt sind (bzw. im Spezialfall, dass keine zufälligen Effekte modelliert sind), heißt der Vektor  $\hat{\beta}$  „best linear unbiased estimator“ (BLUE) von  $\beta$ . Der Vektor  $\hat{\gamma}$  heißt dann „best linear unbiased predictor“ (BLUP) von  $\gamma$ .  $G$  und  $R$  sind in der Praxis zumeist unbekannt und werden nach einer der oben beschriebenen Methoden (wie ML, REML oder MIVQUE0/MINQUE) geschätzt. Die Schätzer  $\hat{\beta}$  und  $\hat{\gamma}$  heißen dann *EBLUE* und *EBLUP*, wobei „E“ für *empirisch* steht, da  $G$  aus den Daten geschätzt wird.

Die Kovarianzmatrix der Schätzfehler ( $\hat{\beta} - \beta, \hat{\gamma} - \gamma$ ) lautet

$$C = \begin{pmatrix} X'R^{-1}X & X'R^{-1}Z \\ Z'R^{-1}X & Z'R^{-1}Z + G^{-1} \end{pmatrix}^{-}, \quad (2.4)$$

deren Komponenten  $G$  und  $R$  dann ebenfalls durch ihre Schätzer ersetzt werden.

### Schätzprobleme

Bei komplexen gemischten Modellen können die weiter oben zitierten Iterationsalgorithmen Konvergenzprobleme in der Weise zeigen, dass die Algorithmen gar nicht bzw. zu Werten konvergieren, die außerhalb des Parameterraums liegen. In solchen Fällen kann häufig durch Wahl von geeigneten Startwerten Abhilfe geschaffen werden.

Eine weitere Quelle von Schätzproblemen kann bedingt durch geringe Streuungen zwischen den Stufen der zufälligen Effekte gegeben sein, insbesondere wenn negative Werte geschätzt werden (siehe auch Seite 67). Bei Einrichtungsvergleichen kann – da Varianzen nicht negativ sein dürfen und so der Schätzfehler minimiert werden kann – hier schlicht 0 als wahre Varianz zwischen den Stufen angenommen werden. Die Nullhypothese, dass keine Unterschiede zwischen den Stufen (hier Einrichtungen) bestehen, kann somit angenommen werden. In anderen Situationen kann dieser Ausweg jedoch nicht angezeigt sein, was eine Reparametrisierung des Modells erforderlich machen kann. Ein Vorschlag hierzu ist bei Verbeke und Molenberghs [57], Kapitel 5.6.1, nachzulesen.

### Inferenz im gemischten Modell

In gemischten Modellen werden die festen und die zufälligen Effekte getrennt hinsichtlich ihres (signifikanten) Einflusses auf die Zielgröße getestet. *Wald-Z-Statistiken*, die auf der Normalverteilungs-Annahme der Effekte beruhen, berechnen sich aus dem Parameterschätzer, dividiert durch den asymptotischen Standardfehler  $\widehat{SE}_{\beta_i}$ . Dieser wird aus der Inversen der zweiten Ableitung der Likelihood-Funktion, nach dem jeweiligen Parameter  $\beta_i$ , gebildet.

Für ein einzelnes Element  $\beta_i$  aus dem gesamten Parametervektor  $\beta$  ist für die Nullhypothese

$$H_0 : \beta_i = \beta_{i0} \quad \text{vs.} \quad \beta_i \neq \beta_{i0}$$

die Statistik

$$Z = \frac{\hat{\beta}_i - \beta_{i0}}{\widehat{SE}_{\beta_i}}$$

asymptotisch standardnormalverteilt. Für beliebige Linearkombinationen  $L$  mit Rang  $l$  gilt unter Nullhypothese

$$H_0 : L\beta = \beta_0 \quad \text{vs.} \quad L\beta \neq \beta_0 :$$

$$Z = (\hat{\beta} - \beta_0)' L \left( L \left( \sum_{i=1}^n X_i' V_i^{-1}(\hat{\gamma}) X_i \right)^{-1} L' \right)^{-1} \sim \chi_l^2 .$$

Die geschätzten Standardfehler von  $\hat{\beta}$  unterschätzen im gemischten Modell allerdings die tatsächliche Variabilität in  $\beta$  (siehe Dempster et al. [9]). Diese Verzerrung kann beispielsweise durch die Verwendung von F- bzw. t-Tests über die Verteilung von  $\beta$  umgangen werden. Die Zahl der Freiheitsgrade muss hier noch z.B. mittels Satterthwaite-Verfahren [48] aus den Daten geschätzt werden. Weitere Methoden, wie natürlich die Likelihood-Ratio-Tests oder Methoden zur Korrektur der Freiheitsgrade in Wald-Statistiken, sind verfügbar. Bei hinreichend großen Stichprobenumfängen, wie es zumeist bei Registern im Gesundheitswesen der Fall ist, werden sich die unterschiedlichen Methoden jedoch hinsichtlich der Ergebnisse kaum auswirken.

Bei Einrichtungsvergleichen ist man ohnehin in der Regel weniger an dem Einflussgrad der festen Effekte, denn an denen der Varianzkomponenten interessiert. Nach der Bestimmung der Schätzer für die Zentrumsmitelwerte ist bei Einrichtungsvergleichen die Bestimmung der Konfidenzintervalle ein zentraler Punkt, da diese in der Öffentlichkeit am ehesten verstanden werden und gleichzeitig auch Testcharakter haben.

Für Tests zur Bestimmung von zufälligen Effekten im gemischten Modell werden Likelihood-basierte Statistiken verwendet. Eine solche ist die *Wald-Z-Statistik*, die auf der Normalverteilungs-Annahme der Effekte beruht.  $Z$  berechnet sich aus dem Parameterschätzer, dividiert durch den geschätzten asymptotischen Standardfehler  $\widehat{\text{SE}}_{\gamma_i}$ , der sich wiederum aus der Inversen der partiellen Ableitung zweiter Ordnung (nach  $\gamma$ ) berechnet.

Für nicht normalverteilte Effekte in  $G$  können auch Likelihood-Ratio-Statistiken verwendet werden (siehe beispielsweise Self und Liang [51]).

### 2.3.2 Modelle mit zufälligen Effekten, Varianzkomponenten-Modelle

Einen Spezialfall gemischter Modelle stellen die Varianzkomponenten-Modelle dar. Hier werden, abgesehen vom Populationsmittel  $\mu$ , ausschließlich zufällige Effekte modelliert.

In Datenerhebungen im Bereich des Gesundheitswesens trifft man häufig eine Situation an, bei der Daten in hierarchischer Form vorliegen, die dann auch „**multilevel data**“ genannt werden. Eine solche Datenlage liegt beispielsweise vor, wenn Patienten einer bestimmten Population – etwa mit einer bestimmten Indikation bzw. Erkrankung – in unterschiedlichen Kliniken behandelt werden. Die Kliniken werden hierbei nicht als feste Einflussgröße, sondern als Repräsentanten einer Zufallsstichprobe, die aus einer theoretisch unendlichen Grundgesamtheit gezogen wurden, aufgefasst. Grundlagen werden in zahlreichen Lehrbüchern als Teil der gemischten Modelle beschrieben. Beispielfhaft seien hier Leyland und Goldstein [34] oder Verbeke und Molenberghs [57] genannt.

Nimmt man ferner an, dass sich die Ausprägungen der Zielgröße  $Y$  (z.B. Behandlungserfolg) zwischen den Zentren systematisch unterscheiden, so folgt, dass die Variabilität innerhalb der Zentren geringer ist als die Variabilität der Gesamtpopulation. In einer solchen Datenlage wäre die niedrigste Hierarchiestufe („level-1“) durch die Patienten (innerhalb der Klinik) besetzt, die zweite Stufe („level-2“) durch die behandelnde Klinik selbst.

Eine hierarchische Datenlage kann ebenfalls durch Messwiederholungen gegeben sein, die bei ein und demselben Patienten durchgeführt werden. In diesem Fall wäre die erste Ebene der Messzeitpunkt, die zweite Ebene der Patient und die dritte Ebene die behandelnde Klinik. Bei Einrichtungsvergleichen und einer Situation mit zwei Ebenen können die Patienten innerhalb der Einrichtungen als Messwiederholungen aufgefasst werden. Hier bezeichnen die Subjekte die Messwiederholungen im Zentrum, und deren Abweichung vom Zentrumsmittelwert wäre der Restfehler des Modells.

Eine solche Situation kann mit einfacher Varianzanalyse (vgl. Kap. 2.2.1.1, „Einfachklassifikation“) ausgewertet werden, bei der die Kliniken als einziger Faktor mit

festen Effekten modelliert werden. Man erhält Schätzer für die jeweiligen Zentrums-  
effekte  $\hat{\beta}_i$  sowie für die Restvarianz  $\sigma_e^2 > 0$ , die zunächst für alle Zentren als identisch  
angenommen wird („Homoskedastizität“). Auf mögliche Unterschiede zwischen den  
Zentren kann dann mittels F- bzw. t-Tests geprüft werden.

### 2.3.2.1 Beispiel: Die einfache hierarchische Klassifikation

Interpretiert man die Versuchseinheiten (also z.B. die zu vergleichenden Einrichtun-  
gen) als Realisation einer Zufallsstichprobe aus einer (theoretisch unendlich großen)  
Grundgesamtheit und will man von diesen Stichprobenelementen auf die Gesamtva-  
riabilität in der Grundgesamtheit schließen, kommt man zu Modellen mit zufälligen  
Effekten (Modelle II der Varianzanalyse).

Hier interessiert man sich nicht primär für den **dann zufälligen** Effekt einer be-  
stimmten Versuchseinheit, sondern eher für das Verhältnis zwischen der Variabilität  
zwischen den Stufen („Signalstärke“) und der Reststreuung (auch als „Grunddrau-  
schen“ bezeichnet). Diese Stufen werden auch „Varianzkomponenten“ genannt. Man  
möchte aufgrund der Realisation der Stichprobe, deren Auswahl als repräsentativ  
angenommen wird, auf die wahre Variation der Effekte der Grundgesamtheit schlie-  
ßen.

*Begründer der Schätzung von Varianzkomponenten sind Friedrich Robert  
Helmert (1876) [23], der ein Verfahren zur Schätzung von Varianzkom-  
ponenten in Zusammenhang mit der Vermessung der Gotthard-Tunnel-  
Achse angibt, und unabhängig davon Ronald Aylmer Fisher, der allge-  
meine Modelle zur Schätzung von Varianzkomponenten in den zwanziger  
Jahren entwickelte.*

Das einfachste hierarchische Modell mit zwei Ebenen kann in der Summations-  
schreibweise wie folgt formuliert werden:

$$y_{ij} = \mu + a_i + e_{ij} \quad \text{mit} \quad i = 1, \dots, p \quad \text{und} \quad j = 1, \dots, n_i .$$

Der Populationsmittelwert  $\mu$  ist als fester Effekt und die  $a_i$  sind als zufällige  
Effekte modelliert, wobei  $a_i$  für alle Patienten *innerhalb* des  $i$ -ten Zentrums festge-  
halten ist; der Term  $e_{ij}$  bezeichnet den (Rest-)Fehler bezüglich des  $j$ -ten Patienten

im  $i$ -ten Zentrum. In der Notation des gemischten Modells schreiben sich  $G$  als Diagonalmatrix mit  $p$  Elementen (ein Faktor mit zufälligen Effekten)  $\sigma_a^2$  und  $R$  als Diagonalmatrix  $\sigma_e^2 I_n$ .

Die beiden Zufallsterme  $\gamma_i$  und  $e_{ij}$  sind jeweils als stochastisch unabhängig normalverteilt mit Erwartungswert 0 und Varianz  $\sigma_a^2$  bzw.  $\sigma_e^2$  angenommen. Somit gilt

$$E(y_{ij}) = \mu \quad \text{für alle } i, j .$$

Die Gesamtvarianz  $\sigma^2$  der Zielgröße zerfällt somit in zwei **unabhängige** Komponenten:

- die Varianz der zufälligen Effekte  $\gamma_i$ ,  $\sigma_a^2$  als Variation **zwischen** den Versuchseinheiten und
- die Varianz  $\sigma_e^2$  **innerhalb** der Versuchseinheiten (Restfehler der Individuen; Streuung der Level-1 Residuen).

Die Varianz-Kovarianzmatrix von  $Y$  schreibt sich zu

$$V = ZGZ' + R \quad \text{mit}$$

$$\text{Var}(y_{ij}) = \text{Var}(\gamma_i) + \text{Var}(e_{ij}) = \sigma^2 = \sigma_a^2 + \sigma_e^2 ,$$

die für alle  $i$  und  $j$  gilt, Homoskedastizität vorausgesetzt.

Gemäß der Notation des gemischten Modells gilt für die als normalverteilt mit Erwartungswert 0 angenommenen Zufallsterme  $\gamma = a$  und  $e$ :

$$E \begin{pmatrix} a \\ e \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

$$D = \text{COV} \begin{pmatrix} a \\ e \end{pmatrix} = \begin{pmatrix} \sigma_a^2 & 0 \\ 0 & \sigma_e^2 \end{pmatrix} ,$$

und für die Beobachtungen

$$\begin{pmatrix} \gamma_i \\ \bar{y}_i. \end{pmatrix} \sim N \left( \begin{pmatrix} 0 \\ \mu \end{pmatrix}, \begin{pmatrix} \sigma_a^2 & \sigma_a^2 \\ \sigma_a^2 & \sigma_a^2 + \frac{\sigma_e^2}{n_i} \end{pmatrix} \right). \quad (2.5)$$

Aufgrund der hierarchischen Struktur des Modells sind somit die Messwiederholungen (bzw. hier die Subjekte) innerhalb der Stufen korreliert, jedoch unabhängig zwischen den Stufen:

$$COV(y_{ij}, y_{i'j'}) = \begin{cases} \sigma_a^2 + \sigma_e^2, & \text{falls } i = i' \text{ und } j = j', \\ \sigma_a^2, & \text{falls } i = i' \text{ und } j \neq j', \\ 0, & \text{sonst.} \end{cases}$$

Die  $n \times n$ -Matrizen  $ZGZ'$  und  $V$  weisen somit eine Blockstruktur entlang der Hauptdiagonalen mit  $p$  ( $n_i \times n_i$ )-Blöcken auf.

Beispiel für  $p = 2$ ,  $n_1 = 2$  und  $n_2 = 3$ :

$$ZGZ' = \begin{pmatrix} \sigma_a^2 & \sigma_a^2 & 0 & 0 & 0 \\ \sigma_a^2 & \sigma_a^2 & 0 & 0 & 0 \\ 0 & 0 & \sigma_a^2 & \sigma_a^2 & \sigma_a^2 \\ 0 & 0 & \sigma_a^2 & \sigma_a^2 & \sigma_a^2 \\ 0 & 0 & \sigma_a^2 & \sigma_a^2 & \sigma_a^2 \end{pmatrix} \quad V = \begin{pmatrix} \sigma^2 & \sigma_a^2 & 0 & 0 & 0 \\ \sigma_a^2 & \sigma^2 & 0 & 0 & 0 \\ 0 & 0 & \sigma^2 & \sigma_a^2 & \sigma_a^2 \\ 0 & 0 & \sigma_a^2 & \sigma^2 & \sigma_a^2 \\ 0 & 0 & \sigma_a^2 & \sigma_a^2 & \sigma^2 \end{pmatrix}$$

Ein lineares Modell, bei dem sich die Gesamtvarianz aus mehreren zufälligen Komponenten zusammensetzt, nennt man VARIANZKOMPONENTEN-MODELL. Wie in der einfachen Varianzanalyse, können die Modell-Koeffizienten über die Gruppenmittelwerte geschätzt werden. Es wird jedoch später gezeigt werden, dass dieser Ansatz nur bei ausschließlich festen Effekten korrekt ist.

Die Analysestafel der einfachen hierarchischen Klassifikation hat (für balancierte Versuche, d.h.  $n_i = q$  für alle  $i$ ) die in Tabelle 2.3 dargestellte Gestalt.

Die Abkürzungen gelten analog zu Tabelle 2.1, wobei

SSA                      Sum of Squares for random effect A (Quadratsumme der Stufen)

bezeichnet.

**Tabelle 2.3: Tafel der einfachen hierarchischen Klassifikation, balancierte Situation**

Variations- ursache	Freiheits- grade	Quadratsummen	Mittlere Quadratsumme
Unterschiede zwischen den Stufen	$p - 1$	$SSA = q \sum_{i=1}^p (\bar{y}_{i.} - \bar{y}_{..})^2$	$MSA = \frac{SSA}{p - 1}$
innerhalb der Stufen	$p(q - 1)$	$SSE = \sum_{i=1}^p \sum_{j=1}^q (y_{ij} - \bar{y}_{i.})^2$	$MSE = \frac{SSE}{p(q - 1)}$
Gesamt	$pq - 1$	$SSG = SSA + SSE$	

Für die Erwartungswerte der Streuungskomponenten gilt:

$$E(MSA) = \sigma_e^2 + q\sigma_a^2 ,$$

$$E(MSE) = \sigma_e^2 .$$

Die Streuungskomponenten  $\sigma_a^2$  und  $\sigma_e^2$  können somit durch

$$\hat{\sigma}_a^2 = s_a^2 = \frac{(MSA - MSE)}{q} \quad \text{und} \quad \hat{\sigma}_e^2 = s_e^2 = MSE$$

erwartungstreu geschätzt werden. Diese Methode wird auch als „Typ-1-Schätzung bezeichnet“ (vgl. Gaylor et al. [15]).

Bei unbalancierten Versuchen, d.h. die variablen Stufenumfänge  $n_i$  ersetzen die feste Wiederholungszahl  $q$ , ergibt sich die in Tabelle 2.4 dargestellte Analysetafel.

Für die Erwartungswerte der Streuungskomponenten gilt jetzt:

$$E(MSA) = \sigma_e^2 + Q\sigma_a^2 \quad \text{mit} \quad Q = \frac{N}{p} - \frac{Var(n_i)}{N} ,$$

$$E(MSE) = \sigma_e^2 .$$

Die Streuungskomponenten  $\sigma_a^2$  und  $\sigma_e^2$  selbst können somit durch

$$\hat{\sigma}_a^2 = s_a^2 = \frac{(MSA - MSE)}{Q} \quad \text{und} \quad \hat{\sigma}_e^2 = s_e^2 = MSE$$

**Tabelle 2.4: Tafel der einfachen hierarchischen Klassifikation, unbalancierte Situation**

Variations- ursache	Freiheits- grade	Quadratsummen	Mittlere Quadratsumme
Unterschiede zwischen den Stufen	$p - 1$	$SSA = \sum_{i=1}^p n_i (\bar{y}_{i.} - \bar{y}_{..})^2$	$MSA = \frac{SSA}{p - 1}$
innerhalb der Stufen	$N - p$	$SSE = \sum_{i=1}^p \sum_{j=1}^q (y_{ij} - \bar{y}_{i.})^2$	$MSE = \frac{SSE}{N - p}$
Gesamt	$N - 1$	$SSG = SSA + SSE$	

wieder erwartungstreu geschätzt werden.

Bemerkung: Im Varianzkomponenten-Modell gilt zwar – wie im gewöhnlichen ANOVA-Modell –  $SSG = SSA + SSE$ , jedoch  $MSG \neq MSA + MSE$ . Hierauf wird später noch eingegangen.

Bei geringen Streuungen zwischen den Stufen im Vergleich zur Restfehlerstreuung könnte  $\hat{\sigma}_a^2$  sowohl in der balancierten als auch der unbalancierten Situation unerwünschterweise auch negative Werte annehmen, was zu den in Kapitel 2.3.1 beschriebenen Schätzproblemen führt. Der generelle Einsatz eines positiv semidefiniten wird daher empfohlen. Alternativ zu ML- oder REML-Methoden kann der (stets positive) Schätzer nach Hartung [21]

$$s_a^2(H) = MSA \frac{q}{1 + q^2}$$

verwendet werden. In gemischten Modellen, bei denen das Hauptaugenmerk auf der Bestimmung von Varianzkomponenten (zufällige Effekte) liegt, wird jedoch zur Schätzung der Modell-Komponenten zumeist die zwar numerisch aufwändige, jedoch mit guten Schätzeigenschaften versehene REML-Methode bevorzugt. Bei negativen Schätzungen  $\hat{\sigma}_a^2$  wird hier der wahre Wert  $\sigma_a^2$  in der Regel als 0 angenommen.

### 2.3.3 Varianzkomponenten-Modelle mit Kovariaten

Lässt man nun weitere zwei Gruppen von Kovariaten als feste Effekte zu, so erweitert sich die Modellgestalt zu

$$y_{ij} = \left( \beta_0 + \sum_{l=1}^a \beta_l x_{lij} + \sum_{m=a+1}^{a+b} \beta_m x_{mi} \right) + (\gamma_{0i} + e_{ij}) .$$

Die  $\beta_l$  stellen hier  $a$  feste Einflussfaktoren der Individuen dar (z.B. anamnestiche Variablen der Patienten innerhalb der Zentren), die  $\beta_m$   $b$  feste Faktoren der Zentren (wie z.B. Bettenzahl oder geographische Lage). Dieses Modell hat nun die Gestalt einer multiplen Regressionsanalyse, erweitert um den zufälligen Effekt (oder Zufallsfehler) des Zentrums, also einen Fehlerterm der zweiten Ebene („level-2 error term“ oder „level-2 residual term“).

### 2.3.4 Modelle mit zufälligen Koeffizienten

Die im vorigen Abschnitt beschriebenen festen Regressionskoeffizienten  $\beta_l$  können auch als zufällige Effekte betrachtet werden. Dem Modell der Form  $y = X\beta + e$  wird dann ein weiterer Zufallsterm hinzugefügt:

$$y_{ij} = \left( \beta_0 + \sum_{l=1}^a \beta_l x_{lij} + \sum_{m=a+1}^{a+b} \beta_m x_{mi} \right) + \left( \gamma_{0i} + \sum_{p=1}^c \gamma_{pj} x_{pij} + e_{ij} \right) .$$

In Modellen dieses Typs, der häufig auch als RANDOM SLOPE MODELS bezeichnet wird, kann Unabhängigkeit zwischen den zufälligen Effekten nicht mehr angenommen werden, was die Schätzung der Kovarianzen nötig macht.

### 2.3.5 Varianzexzess bei gemischten linearen Modellen

#### 2.3.5.1 Motivation, Auswirkung zufälliger Effekte

In Modellen mit zufälligen Effekten muss berücksichtigt werden, dass die Verteilung der *beobachteten* Stufenmittelwerte tatsächlich „breiter“ (d.h. mit größerer Varianz

versehen) ist als die unbekannt *tatsächliche* Verteilung. Es gelten dabei mit der Notation aus Kapitel 2.3.2.1, Homoskedastizität (also  $\sigma_{ei} = \sigma_e$  für alle  $i$ ) und Balanciertheit ( $n_i = q$ ) vorausgesetzt, die folgenden Beziehungen:

$$\text{Var}(\bar{y}_{i.}) = \sigma_a^2 + \frac{\sigma_e^2}{q} \quad (1),$$

$$\text{Var}(y_{ij}) = \sigma_a^2 + \sigma_e^2 \quad (2),$$

$$\text{Var}(\bar{y}_{..}) = \text{Var}\left(\frac{1}{p} \sum_{i=1}^p \bar{y}_{i.}\right) = \frac{1}{p^2} \left(p \left(\sigma_a^2 + \frac{\sigma_e^2}{q}\right)\right) = \frac{1}{p} \left(\sigma_a^2 + \frac{\sigma_e^2}{q}\right) \quad (3).$$

Der zweite Term in der Gleichung (1) wird auch als „Varianzexzess“, „Varianzinflation“ oder „Varianzaufblähung“ der Zentrumsmitteiwerte bezeichnet.

Zur Veranschaulichung dieses Effekts wurden einfache Simulationen nach der Situation der einfachen hierarchischen Klassifikation mit den Notationen aus Tabelle 2.3 durchgeführt. Dabei wurden die Realisationen  $y_{ij}$  von  $p = 48$  Zentren mit jeweils  $q = 20$  Subjekten (also  $N = 960$ ) durch Zufallszahlen mit den folgenden Voraussetzungen erzeugt:

$$\mu = 100;$$

$$\gamma_i \sim N(0, \sigma_a^2) \quad \text{und alternativ hierzu:} \quad (1)$$

$$\gamma_i \sim R(-b, b) \quad \rightarrow \quad E(\gamma_i) = 0, \quad \sigma_a^2 = 2b/12, \quad (2)$$

$$\gamma_i \equiv 0 \quad (\text{keine Unterschiede; Nullhypothese gilt}); \quad (3)$$

$$e_{ij} \sim N(0, \sigma_e^2).$$

Die als zufällig modellierten und generierten Zentrumsmitteiwerte entstammen also einer Normal-, Rechtecks- oder Einpunktverteilung. Für die Streuung der wahren Erwartungswerte zwischen den 20 Stufen wurde für die Alternativhypothese (Fälle (1) und (2)) stets  $\sigma_a^2 = 400$  angenommen und für die Streuung innerhalb der Stufen  $\sigma_e^2$  jeweils  $\sigma_a^2$ ,  $2\sigma_a^2$  und  $4\sigma_a^2$  verwendet. Somit ergeben sich sechs Kombinationen, die jeweils mit 10.000 Simulationsläufen durchgeführt wurden:

Verteilung der $\gamma_i$ ( $\sigma_a^2 = 400$ )		$\sigma_e^2$	Varianz- Exzess	$Var(\bar{y}_{i.})$	$Var(y_{ij})$	$Var(\bar{y}_{..})$
1.		400	20	420	800	8,75
2.	(1) Normal	1.600	80	480	2.000	10,00
3.		6.400	320	720	6.800	15,00
4.	(2) Rechteck	400	20	420	800	8,75
5.	$(b = \sqrt{1200} = 34,64)$	1.600	80	480	2.000	10,00
6.		6.400	320	720	6.800	15,00

Die Simulationsergebnisse sind in Abbildung 2.1 dargestellt.

Es zeigt sich, dass die Auswirkung des Varianzexzesses bei  $\sigma_e^2 = \sigma_a^2$  relativ gering ist. Die Verteilung der beobachteten Zentrumsmitteiwerte bei  $\sigma_e^2 = 4\sigma_a^2$  ist jedoch selbst bei rechtecksverteilten Zentrums-Erwartungswerten praktisch nicht mehr von derjenigen bei normalverteilten Zentren unterscheidbar.

Im Fall (3) existieren keinerlei Unterschiede zwischen den Zentren. Also ist  $\sigma_a^2 = 0$ , und es gilt für die beobachteten Mittelwerte ebenfalls  $E(\bar{y}_{i.}) = \mu$ , jedoch aufgrund der Exzessvarianz

$$Var(\bar{y}_{i.}) = \sigma_a^2 + \frac{\sigma_e^2}{q} = \frac{\sigma_e^2}{q} \quad \text{und}$$

$$Var(y_{ij}) = \sigma_a^2 + \sigma_e^2 = \sigma^2 .$$

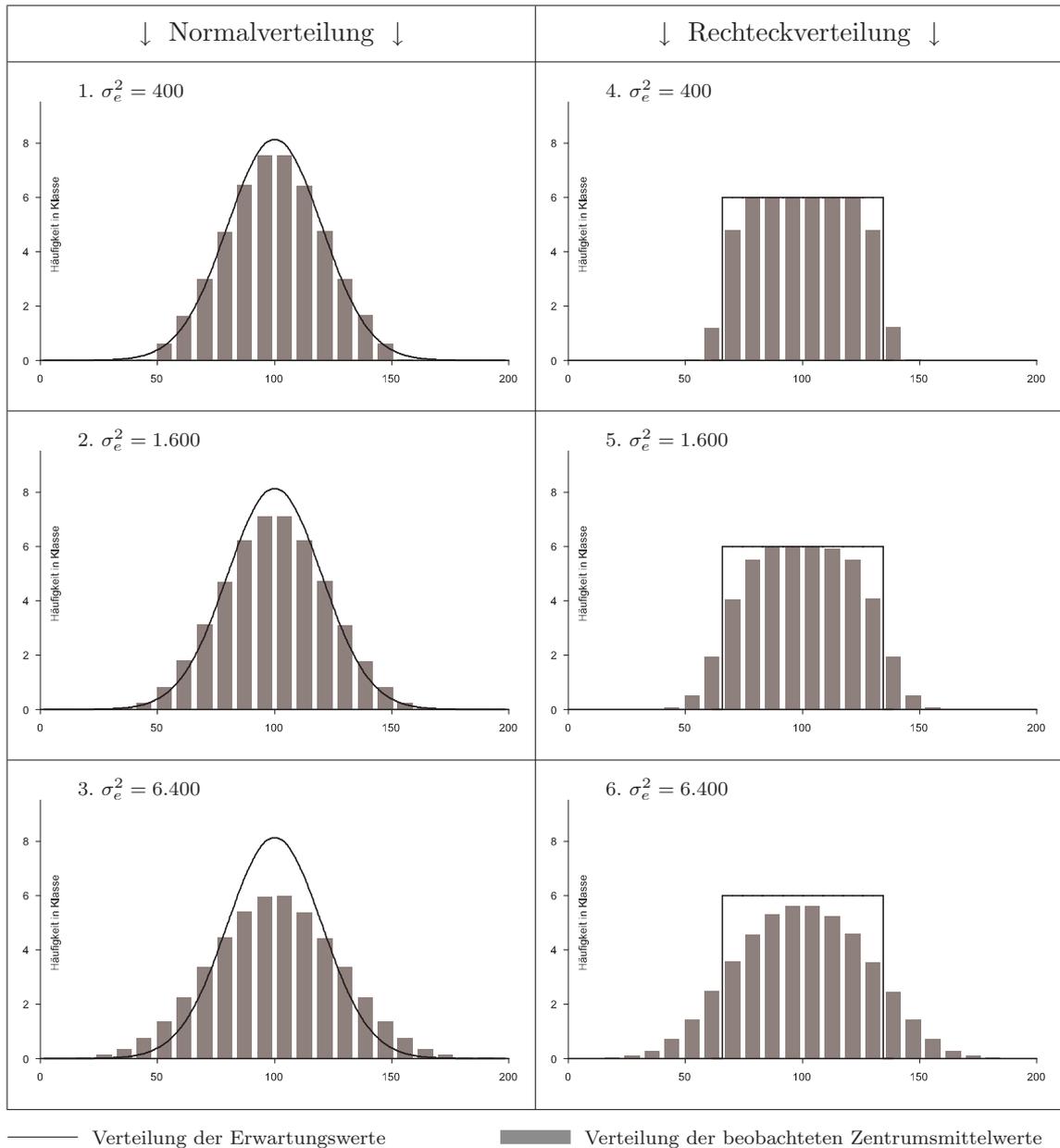
Für den Fall (3) in obigem Beispiel mit  $\sigma_e^2 = 6.400$  zeigt die Simulation die in Abbildung 2.2 dargestellte Verteilung der Zentrumsmitteiwerte.

In der Praxis der Einrichtungsvergleiche (basierend auf Registerdaten) sind die Versuche allerdings in den seltensten Fällen balanciert; mitunter können die Fallzahlen zwischen den Kliniken sehr stark variieren. Für den Varianzexzess des  $i$ -ten Zentrums gilt

$$Var(\bar{y}_{i.}) = \sigma_a^2 + \frac{\sigma_e^2}{n_i} .$$

Vergleicht man die erwarteten Streuungen der Zentrumsmitteiwerte  $\bar{y}_{i.}$  für den unbalancierten Fall  $Var_u(\bar{y}_{i.})$  und für den unbalancierten Fall  $Var_b(\bar{y}_{i.})$  bei gleichen

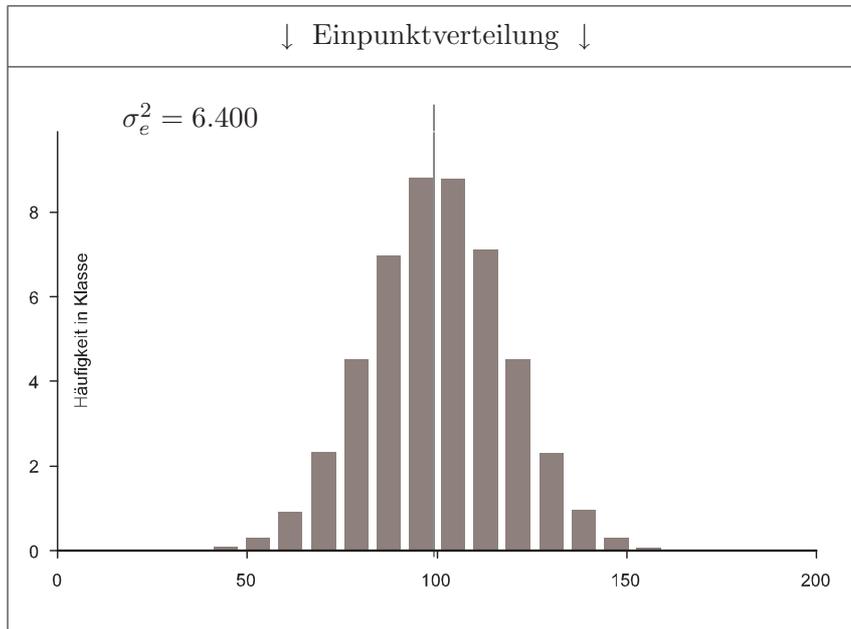
Abbildung 2.1: Simulationsergebnisse zum Varianz-Exzess;  
 Alternativhypothese ( $\sigma_a^2 = 400$ )



Streuungsparametern und gleichen Gesamtfallzahlen miteinander, so gilt

$$\begin{aligned}
 E(Var_u - Var_b) &= E\left(\left(\sigma_a^2 + \frac{\sigma_e^2}{Q}\right) - \left(\sigma_a^2 + \frac{\sigma_e^2}{q}\right)\right) \\
 &= \frac{\sigma_e^2}{q} - \frac{\sigma_e^2}{Q},
 \end{aligned} \tag{2.6}$$

**Abbildung 2.2: Simulationsergebnisse zum Varianz-Exzess;  
Nullhypothese ( $\sigma_a^2 = 0$ )**



wobei  $Q$  als modifizierte Wiederholungszahl in der Form

$$Q = \frac{N^2 - p \operatorname{Var}(n_i)}{Np} = \bar{n}_i - \frac{\operatorname{Var}(n_i)}{N}$$

$$= \frac{1}{N-1} \left( N - \frac{1}{N} \sum_{i=1}^p n_i^2 \right)$$

angegeben werden kann. Bei unbalancierten Versuchen steigt die Streuung der beobachteten Stufenmittelwerte  $\operatorname{Var}(\bar{y}_{i.})$  also im Mittel um den in Gleichung (2.6) dargestellten Wert an. Dieser ist offensichtlich nicht von der Streuung zwischen den Zentren abhängig, sondern nur von der Variation der Fallzahlen, und ist stets  $\geq 0$ , da  $Q \leq q$  gilt. Nur bei balancierten Versuchen ist  $Q = q$ , da hier  $\operatorname{Var}(n_i) = 0$  gilt. Die Varianz der Klinikmittelwerte wird also durch

$$n_i = q = \frac{N}{p}$$

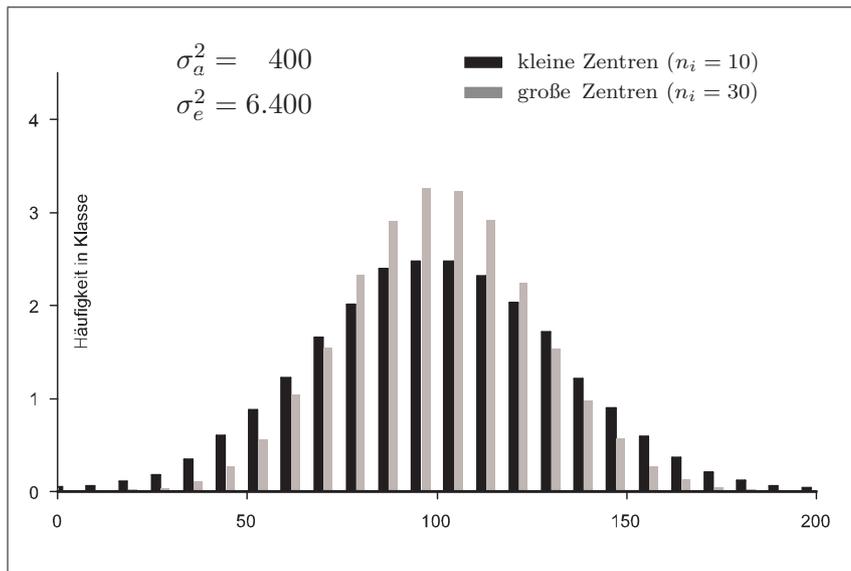
minimiert. Die Gesamtstreuung aller Beobachtungen  $\operatorname{Var}(y_{ij})$  bleibt jedoch unverändert. Der Effekt sei für den oben beschriebenen Fall (3) (also für normalverteilte Zentrumserwartungswerte mit  $\sigma_a^2 = 400$  und  $\sigma_e^2 = 6.400$ ) für zwei Gruppen von Zentrumsgrößen

$$n_i = \begin{cases} 10 & \text{für } i = 1, \dots, 24 \\ 30 & \text{für } i = 25, \dots, 48 \end{cases}$$

anhand von Abbildung 2.3 illustriert. Für den gewählten Fall gilt somit (mit der Notation aus Abschnitt 2.3.2.1):

$$\begin{aligned} q &= 20 \\ Q &= 19,89 \\ E(\text{Var}_u - \text{Var}_b) &= \frac{\sigma_e^2}{q} - \frac{\sigma_e^2}{Q} = 1,7112 . \end{aligned}$$

**Abbildung 2.3: Simulationsergebnisse zum Varianz-Exzess; unbalancierter Fall, normalverteilte Zentrumsmitteiwerte**



Zusätzlich zur Unbalanciertheit wird die Variabilität der Zentren durch möglicherweise vorliegende Heteroskedastizität beeinflusst. Der Varianzexzess des  $i$ -ten Zentrums kann also bei zugelassener Heteroskedastizität im Modell bezeichnet werden als

$$\text{Var}(\bar{y}_i) = \sigma_a^2 + \frac{\sigma_{e_i}^2}{q} \quad \text{bzw.} \quad \text{Var}(\bar{y}_i) = \sigma_a^2 + \frac{\sigma_{e_i}^2}{n_i} .$$

Der Restfehler  $e_{ij}$  entstammt einer Normalverteilung mit Erwartungswert 0 und Klinik-spezifischer Standardabweichung  $\sigma_{e_i}$ , anstelle der gemeinsamen  $\sigma_e$ .

### Weitere Bedeutung von Zufallseffekten

Zusätzlich zu der vergrößerten Varianz wird man bei z.B. jährlich wiederholter Durchführung des Rankings von Einrichtungen beobachten, dass sich Zentren, die bei einer Erhebung einen besonders extremen Rangplatz eingenommen haben, zu späteren Zeitpunkten eher im „Mittelfeld“ befinden werden. Dies ist durch den Umstand begründet, dass die Einflüsse der Zentren zufälligen Schwankungen unterliegen. Ein extremes Ergebnis kann also auch zufällig und muss nicht durch systematische Einflüsse entstehen. Dieser Effekt, der auch in der (belebten und unbelebten) Natur zu beobachten ist, wird häufig als „Regression zum Mittelwert“ („regression to the mean“) bezeichnet und wird besonders bei Einrichtungen mit kleineren Fallzahlen zu beobachten sein.

#### 2.3.5.2 Best Linear Unbiased Prediction – Minimierung des Prognosefehlers

Wie im vorigen Abschnitt beschrieben wurde, hätte eine Benutzung der beobachteten Zentrumsmittelwerte  $\hat{a}_i = \bar{y}_i$  – wie in der Varianzanalyse vorgeschlagen – als Maximum-Likelihood-Schätzung der Erwartungswerte (welche aufgrund der Symmetrie der Verteilung identisch ist mit der KQ-Schätzung) zur Folge, dass die Effekte  $\gamma_i$  überschätzt werden. Die beobachteten Zentrumseffekte schätzen die  $\gamma_i$  zwar erwartungstreu, minimieren also den mittleren quadratischen Fehler innerhalb der Zentren ( $MSE_i$ ),

$$MSE_i = Var(\bar{y}_i) = E((\bar{y}_i - \gamma_i)^2) = \frac{\sigma_e^2}{n_i}, \quad (2.7)$$

jedoch nicht den Gesamtfehler des Versuches MSE,

$$MSE = \sum_{i=1}^p MSE_i = E\left(\sum_{i=1}^p (\bar{y}_i - \gamma_i)^2\right), \quad (2.8)$$

wenn zufällige Effekte vorliegen. Die Verwendung der beobachtete Klassenmittelwerte als konventionelle ANOVA-Schätzer und -Konfidenzintervalle würde in diesem Fall

anti-konservativ wirken, also zu leicht auf Unterschiede zwischen den Stufen schließen lassen. Somit ist es notwendig, die Variation der Klinikmittelwerte entsprechend des Varianzexzesses zu reduzieren und die Schätzer zu modifizieren. Die erwartungstreue Schätzung der Variation der Erwartungswerte zwischen den Zentren ist bereits durch  $\hat{\sigma}_a^2$  gegeben.

Ein Kriterium, welches zur Reduktion der Schätzvarianz verwendet werden kann, ist der Quotient

$$\varrho = \frac{\sigma_a^2}{\sigma_a^2 + \sigma_e^2}, \quad (2.9)$$

der allgemein als **Intraklassen-Korrelationskoeffizient** bezeichnet wird.  $\varrho$  misst die Reliabilität (Zuverlässigkeit) des Ergebnisses, also die Abhängigkeit zwischen den Beobachtungen  $y_{ij}$  und  $y_{ij'}$  ( $j \neq j'$ ) innerhalb desselben Clusters  $i$  und liegt theoretisch im Intervall  $[0, 1]$ . Der Wert von  $\varrho$ , der auch als „Signal-Rauschen-Verhältnis“ („signal-noise ratio“) bezeichnet wird, gibt somit an, wie groß der Anteil der Streuung, die auf dem Effekt der Kliniken beruht, im Verhältnis zur gesamten Streuung ist.

Der Quotient

$$\varrho_i = \frac{\sigma_a^2}{\sigma_a^2 + \frac{\sigma_e^2}{n_i}} \quad (2.10)$$

heißt **Wiederholbarkeit** („reproducibility“) und gibt an, wie groß der Anteil der Signalstärke an der Streuung der Mittelwerte ist.

Unter Verwendung von Gleichung (2.5) lösen die Schätzer (siehe etwa Kackar und Harville [30] oder Verbeke und Molenberghs [57], Kapitel 2.3.2)

$$\begin{aligned} \hat{\mu}_i^{\text{EBLUP}} &= \hat{\mu} + E(\gamma_i | \bar{y}_i) \\ &= \hat{\mu} + E(\gamma_i) + \frac{\text{Cov}(\gamma_i, \bar{y}_i)}{\text{Var}(\bar{y}_i)} (\bar{y}_i - E(\bar{y}_i)) \\ &= \hat{\mu} + \hat{\varrho}_i (\bar{y}_i - \hat{\mu}) \\ &= \hat{\mu} + \frac{n \sigma_a^2}{n \sigma_a^2 + \sigma_e^2} (\bar{y}_i - \hat{\mu}) \end{aligned} \quad (2.11)$$

das in (2.8) gegebene Minimierungsproblem und erfüllen somit die EBLUP-Eigenschaft. Unter weiterer Verwendung der in Kapitel 2.3.2.1 beschriebenen Schätzer  $s_a^2$  und  $s_e^2$

für die Streuungskomponenten bzw.  $\hat{\beta}$  für die festen Effekte können diese somit zur Korrektur des Varianzexzesses benutzt werden.

Diese Schätzung kann auch über Ansätze der Bayesianischen Statistik interpretiert werden („Empirical Bayes Schätzer (EBS)“). Die beobachtete (posteriori-) Verteilung der zufälligen Effekte wird mit ihrer (a-priori-)Verteilung, d.h. bevor Kenntnis über ihre Streuung erlangt wird, gewichtet werden. Der Zusammenhang zwischen den Empirical-Bayes-Schätzern und dem EBLUP-Ansatz wird etwa von Robinson [47] diskutiert.

Nach der beschriebenen Methode liegen somit alle EBLUP-Schätzer zwischen dem beobachteten Zentrumsmittelwert und dem Gesamtmittelwert, da  $\varrho_i \geq 0$  ist. Die Streuung der EBLUP-Schätzer ist also geringer als die der beobachteten Mittelwerte. Dieser Effekt wird allgemein als „Schrumpfung“ oder „shrinkage“ bezeichnet.

Die Schrumpfung der beobachteten Stufenmittelwerte, also die Verschiebung hin zum Gesamtmittel, fällt umso größer aus,

- je kleiner die Fallzahl  $n_i$  des Zentrums,
- je größer der beobachtete Effekt (absolute Differenz zum Gesamtmittel),
- je größer die Streuung innerhalb der Zentren (geschätzte Restfehlerstreuung (MSE)), bzw.
- je kleiner die Streuung zwischen den Zentren (MSA) ist.

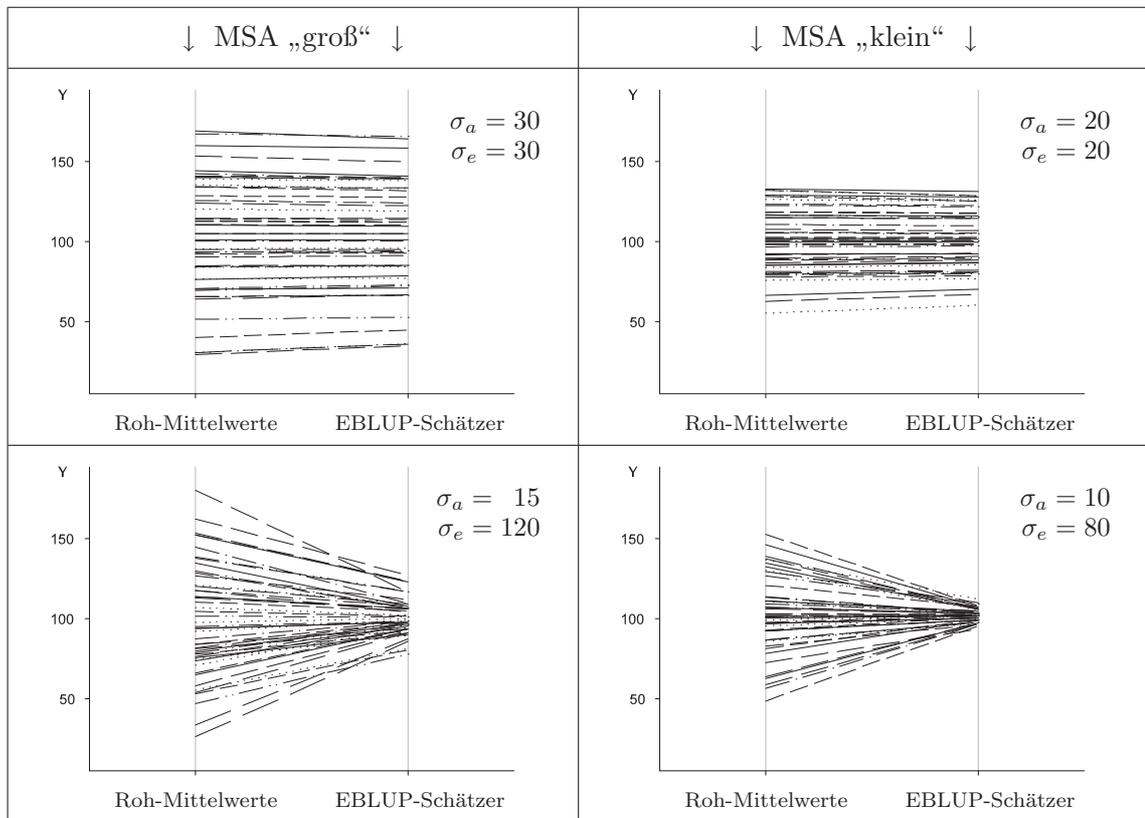
Zentren mit derselben Fallzahl erfahren bei angenommener Homoskedastizität, relativ zur Abweichung des Zentrumsmittelwerts vom Gesamtmittel, dieselbe Korrektur  $\varrho_i$ , da

$$n_i = n_{i'} \quad \Rightarrow \quad \varrho_i = \varrho_{i'}$$

gilt. Abbildung 2.4 zeigt beispielhaft die Wirkung der Schrumpfung für verschiedene Fälle, bei geringer bzw. hoher Restfehlerstreuung im Vergleich zur Streuung zwischen den Zentren.

Die Schrumpfungswirkung ist bei hoher Restfehlerstreuung (geringes „Signal“, hohes „Rauschen“) stärker ausgeprägt, als wenn die Restfehlerstreuung in etwa gleich der Streuung zwischen den Zentren ist.

**Abbildung 2.4: Auswirkung der Schrumpfung auf Effekt-Schätzer**  
 $N = 960, p = 48$ , unbalancierter Fall



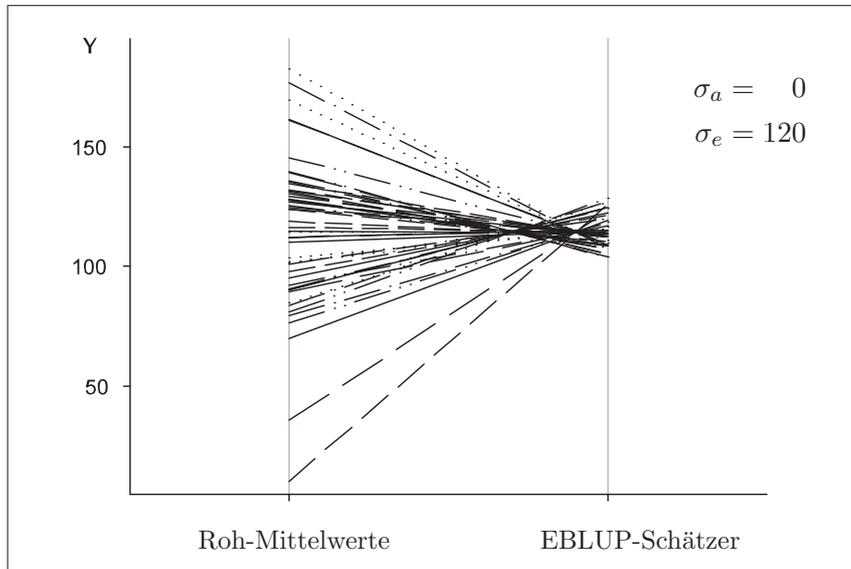
Wegen der Imbalance des Versuchs ist zusätzlich zu beobachten, dass ein Zentrum  $i$ , das sich auf Basis der Mittelwerte besser (oder schlechter) als ein anderes Zentrum  $i'$  darstellt, nach der Korrektur tatsächlich schlechter (oder besser) eingestuft werden kann. In diesem Fall, der jedoch nur für Paare von Zentren mit ungleichen Fallzahlen ( $n_i \neq n_{i'}$ ) möglich ist, schneiden sich die Linien in Abbildung 2.4.

Wie in Kapitel 2.3.2.1 beschrieben, kann  $\hat{\sigma}_a^2$  insbesondere bei schwachen oder nicht existenten Klinik-Effekten negative Werte annehmen, was dazu führt, dass die Kriterien  $\varrho$  und  $\varrho_i$  negativ werden. Eine solche Situation, die bei  $\sigma_a = 0$  tatsächlich mit 50% Wahrscheinlichkeit eintritt, hätte zur Folge, dass die EBLUP-Schätzer aller Zentren in einen – von den beobachteten Mittelwerten aus betrachtet – Bereich jenseits des geschätzten Gesamtmittels klassifiziert würden (siehe Abbildung 2.5).

Dieser Fall sollte jedoch aus mehreren Gründen ausgeschlossen werden,

- aus methodischer Sicht, weil negative Varianzen unerwünscht sind und weil

Abbildung 2.5: Auswirkung negativer Schätzung der Zentrums-Variabilität



negative Korrelationen in der vorliegenden Fragestellung nicht interpretierbar sind, sowie

- aus inhaltlicher Sicht, weil ein Einrichtungsvergleich, bei dem nach Betrachtung von Roh-Mittelwerten „schlechtere“ Zentren als besser eingestuft werden als „bessere“, zurecht kaum öffentliche Akzeptanz finden würde.

Daher sollte für  $\sigma_a^2$  stets ein (zumindest schwach) positiv definierter Schätzer benutzt werden (vgl. Kapitel 2.3.2.1, Seite 67). Würde in der in Abbildung 2.5 dargestellten Beispielsituation als Schätzer der Wert 0 verwendet, würden – da dann  $\hat{\rho}_i = 0$  – die EBLUP-Schätzer aller Zentren den Wert des geschätzten Gesamtmittels  $\hat{\mu}$  annehmen.

Für die Varianz der Mittelwerte und der Prediktionen gilt im Falle bekannter Parameter:

$$\begin{aligned} \text{Var}(\bar{y}_{i\cdot}) &= \sigma_a^2 + \frac{\sigma_e^2}{n_i} = \frac{\sigma_a^2}{\rho_i} > \sigma_a^2, \\ \text{Var}(\hat{\mu}_i^{\text{BLUP}}) &= \sigma_a^2 * \frac{\sigma_a^2}{\sigma_a^2 + \frac{\sigma_e^2}{n_i}} = \rho_i \sigma_a^2 < \sigma_a^2, \end{aligned}$$

und somit

$$\text{Var}(\hat{\mu}_i^{\text{BLUP}}) < \sigma_a^2 < \text{Var}(\bar{y}_{i\cdot}).$$

Dies gilt analog für die EBLUP-Schätzungen. Allerdings muss für diese berücksichtigt werden, dass die Standardfehler aufgrund der nötigen Schätzung für die unbekanntem Koeffizienten  $\beta$  der festen Effekte in  $X$  größer sind als im Falle bekannter Parameter. Damit ist die Verteilung der (empirischen) Prediktoren im Mittel um den Faktor  $\varrho_i$  enger als die tatsächliche unbekannte Verteilung der Erwartungswerte.

Shen und Louis [52] schlugen vor, für die Schrumpfung der ANOVA-Schätzer anstelle von

$$\begin{aligned} \hat{\mu}_i^{\text{BLUP}} &= \mu + \varrho_i (\bar{y}_i. - \mu) \\ \text{die Korrektur } \hat{\mu}_i^{\text{TG}} &= \mu + \sqrt{\varrho_i} (\bar{y}_i. - \mu) \end{aligned} \quad (2.12)$$

zu verwenden und nannten diesen Ansatz „Triple Goal“-Schätzung (kurz „TG-Schätzung“). Im Falle unbekannter Parameter wird wieder  $\varrho_i$  durch  $\hat{\varrho}_i$  ersetzt. Für die (empirischen) Triple-Goal-Schätzer ((E)TG-Schätzer) gilt bei Balanciertheit

$$\begin{aligned} \text{Var}(\hat{\mu}_i^{\text{TG}}) &= \varrho_i \text{Var}(\bar{y}_i.) = \sigma_a^2 \quad \text{bzw.} \\ \text{Var}(\hat{\mu}_i^{\text{ETG}}) &= \hat{\varrho}_i \text{Var}(\bar{y}_i.) = \hat{\sigma}_a^2. \end{aligned}$$

Bei nicht balancierten Versuchen gelten die Beziehungen zumindest approximativ. Die (E)TG-Schätzer reflektieren somit die Verteilung der wahren zufälligen Effekte am besten, obwohl sie nicht den mittleren Prognosefehler minimieren.

### Einrichtungsvergleiche auf Basis der EBLUP-Schätzer

Zunächst werden auf Basis der REML-Methode die festen Effekte und die Varianzkomponenten geschätzt. Die Lösung für die Maximum-Likelihood-Schätzung des Populationsmittelwerts lässt sich, abhängig von  $V$ , wie folgt gewinnen (vgl. Searle [50], Kapitel 3.3):

$$\hat{\mu}^{\text{ML}} = \frac{\sum_{i=1}^p \frac{\bar{y}_i.}{\text{Var}(\bar{y}_i.)}}{\sum_{i=1}^p \frac{1}{\text{Var}(\bar{y}_i.)}} = \frac{\sum_{i=1}^p \frac{\sum_{j=1}^{n_i} y_{ij}}{\sigma_e^2 + n_i \sigma_a^2}}{\sum_{i=1}^p \frac{n}{\sigma_e^2 + n_i \sigma_a^2}}.$$

Bei unbalancierten Versuchen ist somit das Prediktionsmittel nicht gleich dem beobachteten Gesamtmittel über die Zentren. Um zu prüfen, ob tatsächlich Unterschiede zwischen den Zentren bestehen, wird die Nullhypothese

$$H_0 : \sigma_a^2 = 0 \quad \text{vs.} \quad H_1 : \sigma_a^2 \neq 0$$

mittels Wald-Tests geprüft, wobei der Schätzer  $\hat{\sigma}_a^2$  als asymptotisch normalverteilt angenommen wird. Für den Fall, dass die Nullhypothese zum Niveau  $\alpha$  verworfen wird, kann also auf Unterschiede zwischen Kliniken (im Allgemeinen) geschlossen werden.

Für die EBLUP-Schätzer können nun, um zu prüfen, *welche* Kliniken tatsächlich einen Effekt verschieden von 0 aufweisen, Konfidenzintervalle für die  $\hat{\gamma}_i$  angegeben werden. Der Standardfehler wird aus der quadratischen Form der Kovarianzmatrix  $C$  für die Parametervektoren  $\beta$  (feste Effekte) und  $\gamma$  (zufällige Effekte)

$$C = \text{Var}(\hat{\beta} - \beta, \hat{\gamma} - \gamma)$$

bestimmt. Im Falle der einfachen hierarchischen Klassifikation ist  $\beta = \mu$  als einziger fester Effekt und  $\gamma_i$  als Zufallseffekte modelliert;  $C$  ist somit eine  $(p + 1 \times p + 1)$ -Matrix.

### 2.3.5.3 Empirische Eigenschaften von Modell-Schätzern

Die Eigenschaften der Modellschätzer und der EBLUP-Schätzungen für die Zentrumsmittelwerte sollen nun anhand von Simulationen in der einfachen hierarchischen Klassifikation betrachtet werden. Hierzu wurden dieselben Situationen, wie in Kapitel 2.3.5.1 (Fall (2)) verwendet, zugrunde gelegt ( $p = 48$ ,  $q = 20$ ,  $N = 960$ , mit normalverteilten Zentrumsmittelwerten). Es gelte:

$$\begin{aligned} \mu &= 100, \\ \gamma_i &\sim N(0, \sigma_a^2) \quad \text{mit } \sigma_a^2 = 400, \\ e_{ij} &\sim N(0, \sigma_e^2) \quad \text{mit } \sigma_e^2 = 1.600. \end{aligned}$$

Für die Intraklassen-Korrelationskoeffizienten gilt somit

$$\varrho = \frac{\sigma_a^2}{\sigma_a^2 + \sigma_e^2} = \frac{400}{2.000} = 0,20 ,$$

$$\varrho_i = \frac{\sigma_a^2}{\sigma_a^2 + \frac{\sigma_e^2}{n_i}} = \frac{\sigma_a^2}{\text{Var}(\bar{y}_{i.})} = \frac{400}{480} = 0,83 .$$

Für die Simulation wurde unter Verwendung der Software PROC MIXED (SAS<sup>®</sup> Institute, Version 9.1) mittels REML-Methoden Schätzungen für  $\sigma_a^2$ ,  $\sigma_e^2$ ,  $X\hat{\beta} = \hat{\mu}$  und  $Z\hat{\gamma} = \hat{\gamma}_i$  ( $i = 1, \dots, 48$ ) bestimmt. Die Prediktoren für die Zentrumsmitte werte lauten

$$\hat{\mu}_i^{\text{EBLUP}} = X\hat{\beta} + Z\hat{\gamma} = \hat{\mu} + \hat{\gamma}_i .$$

Wie bereits eingangs des Kapitels bemerkt wurde, gilt im Varianzkomponenten-Modell im Gegensatz zum gewöhnlichen ANOVA-Modell mit ausschließlich festen Effekten für die totale Varianz

$$\text{MSG} = \frac{\text{SSG}}{N-1} \neq \text{MSA} + \text{MSE} .$$

Im balancierten Fall gilt aufgrund der Beziehung

$$\sigma_e^2 + q\sigma_a^2 = \frac{\sigma_e^2}{1 - \varrho_i} \quad \Rightarrow \quad E(\text{MSA}) = \frac{E(\text{MSA})}{1 - \varrho_i} :$$

$$\begin{aligned} E(\text{MSG}) &= \frac{E(\text{SSA}) + E(\text{SSE})}{N-1} = \frac{(p-1)(\sigma_e^2 + q\sigma_a^2) + p(q-1)\sigma_e^2}{N-1} \\ &= \sigma_e^2 + \sigma_a^2 \left( \frac{p-1}{p} \frac{N}{N-1} \right) = \sigma_e^2 + \sigma_a^2 \frac{q(p-1)}{N-1} \\ &= \sigma_e^2 + \varrho_i \text{Var}(\hat{\mu}) \left( \frac{N(p-1)}{N-1} \right) \end{aligned} \quad (2.13)$$

und ist damit stets  $< \sigma_a^2 + \sigma_e^2$ , also im gewählten Beispiel = 1.992,1. Für „große“ Zentren tendiert dieser Wert somit gegen

$$\text{MSG}(Y) \approx \sigma_e^2 + \sigma_a^2 \left( \frac{p-1}{p} \right) .$$

Im unbalancierten Fall gilt

$$\begin{aligned}
 E(\text{MSG}) &= \frac{E(\text{SSA}) + E(\text{SSE})}{N-1} = \frac{(p-1)(\sigma_e^2 + Q\sigma_a^2) + (N-p)\sigma_e^2}{N-1} \\
 &= \sigma_e^2 + \sigma_a^2 \left( \frac{p-1}{N/Q} \frac{N}{N-1} \right) = \sigma_e^2 + \sigma_a^2 \frac{Q(p-1)}{N-1} \\
 &= \sigma_e^2 + \frac{\sum \varrho_i \text{Var}(\bar{y}_{i.}) n_i}{N} \left( \frac{Q(p-1)}{N-1} \right)
 \end{aligned} \tag{2.14}$$

und für „große“ Zentren

$$\text{MSG}(Y) \approx \left( \frac{p-1}{N/Q} \right) \sigma_a^2 + \sigma_e^2.$$

Tabelle 2.5 zeigt die Simulationsergebnisse für die oben beschriebene Beispielsituation (100.000 Simulationsläufe).

**Tabelle 2.5: Simulationsergebnisse zu Eigenschaften der EBLUP-Schätzer, balancierter Fall** ( $\sigma_a^2 = 400$ ,  $\sigma_e^2 = 1.600$ ,  $p = 48$ ,  $q = 20$ )

Kenngröße aus Modell		erwartet	Schätzmittel	Schätzmedian
$\hat{\sigma}_a^2$	$= (\text{MSA}-\text{MSE})/q$	400,00	399,78	393,17
$\hat{\sigma}_e^2$	$= \text{MSE}$	1.600,00	1.600,13	1.598,60
$\hat{\mu}$	$= \bar{y}_{..}$	100,00	100,01	100,01
Streuung Kliniken/Patienten		erwartet	beob. Mittel	beob. Median
$\text{Var}(y_{ij})$	$= \text{MSG} (2.13)$	1.992,08	1.991,99	1.986,56
$\text{Var}(\bar{y}_{i.})$	$= \text{MSA}/q = \sigma_a^2 + \sigma_e^2/q$	480,00	479,79	472,81
$\text{Var}(\hat{\mu}_i^{\text{EBLUP}})$	$= \varrho_i \sigma_a^2$	333,33	333,74	326,41
$\text{Var}(\hat{\mu}_i^{\text{ETG}})$	$= \varrho_i \text{Var}(\bar{y}_{i.})$	400,00	399,78	393,17
Streuung Gesamtmittel		erwartet	beobachtet	
$\text{Var}(\bar{y}_{..})$	$= \text{Var}(\bar{y}_{i.})/p$	10,00	9,98	

Wie in Abschnitt 2.3.2.1 ausgeführt wurde, kann – insbesondere in Situationen mit kleiner Gesamtfallzahl oder kleinem Signal-Rauschen-Verhältnis  $\varrho$  – der Fall eintreten, dass eine erwartungstreue (Typ-1-) Schätzung für  $\sigma_a^2$  negativ wird. Bei (Restricted) Maximum-Likelihood-Verfahren ist  $\hat{\sigma}_a^2$  jedoch auf nicht-negative Werte beschränkt. Hieraus ergibt sich ein zu hohes Schätz-Mittel, da im Falle einer negativer (Typ-1-)Schätzung  $\hat{\sigma}_a^2$  auf den Wert Null gesetzt wird. Bei positiven Werten

sind die REML- und Typ-1-Schätzer in der einfach hierarchischen Klassifikation identisch. Bei negativen Werten für  $\hat{\sigma}_a^2$  wird die Schätzung für  $\sigma_e^2$  um denselben Faktor, wie in Gleichung (2.13) dargestellt, modifiziert

$$\hat{\sigma}_e^2 \text{ REML} = \hat{\sigma}_e^2 \text{ Typ1} + \hat{\sigma}_a^2 \text{ Typ1} \left( \frac{p-1}{p} \frac{N}{N-1} \right),$$

und führt dann wieder im Durchschnitt zu geringeren Werten als  $\sigma_e^2$ .

Als Beispiel sei derselbe Fall wie in Tabelle 2.5 betrachtet, jedoch mit kleinerer Dimension,  $p = 4$  und  $q = 5$ , gewählt. Für diese Situation (mit 10.000 Simulationenläufen) lieferten die Typ-1-Schätzungen im Mittel  $\bar{\hat{\sigma}}_a^2 \approx \sigma_a^2$  bzw.  $\bar{\hat{\sigma}}_e^2 \approx \sigma_e^2$ , wobei 27,7% der Typ-1-Schätzungen für  $\sigma_a^2$  negativ waren. Die REML-Schätzungen lieferten im Durchschnitt den Wert  $\bar{\hat{\sigma}}_a^2 \approx 440$  und  $\bar{\hat{\sigma}}_e^2 \approx 1573$ .

Die REML-Methodik führt somit zu einer Überschätzung der Streuung der zufälligen Effekte und einer Unterschätzung des Restfehlers  $\sigma_e^2$  im Durchschnitt; allerdings können die Ergebnisse negativer Varianzkomponenten-Schätzungen für Klinik-Rankings ohnehin nicht verwendet werden. Daher hat dieser Umstand für die korrekte Interpretation der Ergebnisse keine Bedeutung.

Zur weiteren Illustration der Eigenschaften der Klinik-spezifischen EBLUP-Schätzer seien folgende Kenngrößen bezeichnet:

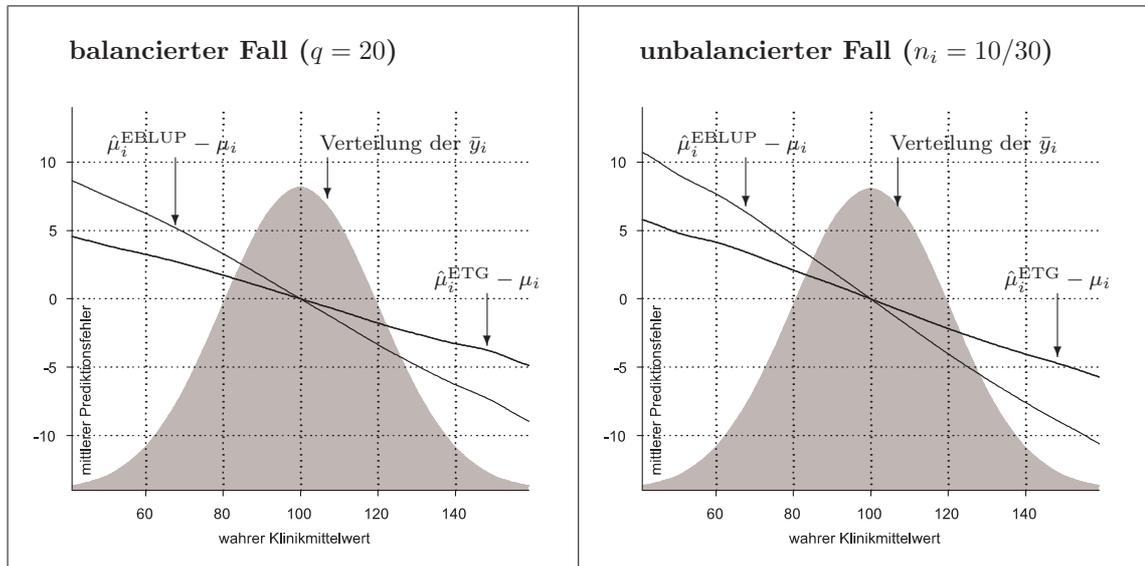
- |     |              |                                |                              |
|-----|--------------|--------------------------------|------------------------------|
| (1) | Schätzer     | des Populationsmittels         | $\hat{\mu}$                  |
| (2) | Schätzfehler | des Populationsmittels         | $\hat{\mu} - \mu$            |
| (3) | Schätzer     | des $i$ -ten Klinikeffekts     | $\hat{\gamma}_i$             |
| (4) | Schätzfehler | des $i$ -ten Klinikeffekts     | $\hat{\gamma}_i - \gamma_i$  |
| (5) | Prediktor    | für Klinik $i$ , $\hat{\mu}_i$ | $\hat{\mu} + \hat{\gamma}_i$ |
| (6) | Schätzfehler | des Prediktors $i$             | $\hat{\mu}_i - \mu_i$ .      |

In Abbildung 2.6 sind die empirischen Schätzfehler der Prediktoren und der Triple-Goal-Schätzer sowie deren Standardabweichungen, in Abhängigkeit vom wahren Zentrumsmittelwert, für die Beispielsituation dargestellt.

Bemerkung: Für die genannten Schätzfehler gilt die folgende Beziehung:

$$\begin{aligned} (4) &= \hat{\gamma}_i - \gamma_i = (\hat{\mu}_i - \hat{\mu}) - (\mu_i - \mu) \\ \Rightarrow (4) - (6) &= \hat{\mu} - \mu. \end{aligned}$$

**Abbildung 2.6: Mittlere Prediktionsfehler bei normalverteilten Mittelwerten**  
 $(\sigma_a^2 = 400, \sigma_e^2 = 1.600, p = 48, N = 960)$



Aufgrund der Erwartungstreue von  $\hat{\mu}$  ist diese Differenz im Mittel gleich 0. Die Standardfehler der Effektschätzer ( $\hat{\gamma}_i - \gamma_i$ ) sind jedoch größer als die Standardfehler der Prediktionen ( $\hat{\mu}_i - \mu_i$ ), da in letztere die Kovarianzen aus der Matrix  $C$  eingehen.

Für den (realistischeren) unbalancierten Fall sei die obige Simulation entsprechend wiederholt. Hierbei wurde die Zahl der Zentren sowie die Gesamtfallzahl festgehalten. Die  $n_i$  wurden wie im Beispiel in Abbildung 2.3 gewählt ( $n_i = 10$  für  $i = 1, \dots, 24$ ;  $n_i = 30$  für  $i = 25, \dots, 48$ ).

Wie beispielsweise von Kacker und Harville [31] beschrieben wurde, wird die tatsächliche Matrix  $C$  von  $\hat{C}$  leicht unterschätzt, da keine Adjustierung hinsichtlich der Ungewissheit von  $G$  und  $R$  berücksichtigt wurde. Diese Unterschätzung wird von der PROC MIXED Software durch die Angabe der approximativen t- bzw. F-Statistiken ausgeglichen. Wie sich in der Simulation zeigt, sind bei hinreichend großer Fallzahl kaum Unterschätzungen von  $\sigma_a^2$  und  $\sigma_e^2$  zu beobachten. Bei kleinen Fallzahlen empfiehlt sich die Verwendung von geeigneter Freiheitsgradadjustierung der REML-Schätzer für die zufälligen Effekte (siehe z.B. Kenward und Roger [32] und Satterthwaite [48] durch Wahl der entsprechenden Option (DDFM=KENWARDROGER)).

**Tabelle 2.6: Simulationsergebnisse zu Eigenschaften der EBLUP-Schätzer, unbalancierter Fall ( $\sigma_a^2 = 400$ ,  $\sigma_e^2 = 1.600$ ,  $p = 48$ ,  $Q = 19, 89$ )**

Kenngröße aus Modell		erwartet	Schätzmittel	Schätzmedian
$\hat{\sigma}_a^2$	$= (\text{MSA-MSE})/Q$	400,00	399,57	392,53
$\hat{\sigma}_e^2$	$= \text{MSE}$	1.600,00	1.599,23	1.597,94
$\hat{\mu}'$	$= \sum \bar{y}_i./p$	100,00	99,98	100,00
$\hat{\mu}^{\text{REML}}$	$= (2.11)$	100,00	99,98	99,98
Streuung Kliniken/Patienten		erwartet	beob. Mittel	beob. Median
$\text{Var}(y_{ij})$	$= \text{MSG (2.14)}$	1.989,99	1.988,81	1.982,78
$\text{Var}(\bar{y}_i.)$	$= \sum(\sigma_a^2 + \sigma_e^2/n_i)/p$	506,67	506,36	499,05
$\text{Var}(\hat{\mu}_i^{\text{EBLUP}})$	$= \sum(\varrho_i \sigma_a^2)/p$	319,33	319,97	312,24
$\text{Var}(\hat{\mu}_i^{\text{ETG}})$	$\approx \sigma_a^2$	400,00	399,66	392,73
Streuung Gesamtmittel		erwartet	beobachtet	
$\text{Var}(\hat{\mu}')$	$= \text{Var}(\bar{y}_i.)/p$	10,56	10,57	

#### 2.3.5.4 Betrachtungen zur Testgüte

Abschließend ist zu prüfen, inwieweit die verwendeten Tests das Signifikanzniveau einhalten. Hierzu wird zunächst – wieder im balancierten und unbalancierten Fall – der Fall  $H_0 : \sigma_a^2 = 0$  betrachtet. In diesem Fall sollte die globale Hypothese (Wald-Z-Test) – unabhängig von der tatsächlichen Restfehlerstreuung  $\sigma_e^2$  – in höchstens  $\alpha/2$  der Versuche abgelehnt werden (siehe Tabelle 2.7, basierend auf 10.000 Simulationsläufen je Kombination), obwohl bei der REML-Methode – wie besprochen – keine erwartungstreue Schätzung möglich ist. Je Kombination wurde jeweils der balancierte und der unbalancierte Fall betrachtet. Im unbalancierten Fall wurden die Zentrumsgrößen bei jeweils gleicher Gesamtfallzahl wie im obigen Beispiel ( $n^{\text{„groß“}} = 3 * n^{\text{„klein“}}$ ) gewählt.

In allen gewählten Situationen waren die Testentscheidungen konservativ, d.h. die Nullhypothese wurde – insbesondere bei wenigen und „großen“ Zentren – in deutlich weniger als 2,5% der Versuche verworfen. In bestimmten Situationen – etwa bei sehr großer Zentren-Anzahl  $p$  und sehr kleiner Wiederholungszahl  $q$  – kann jedoch der Test auch zu häufig auf Unterschiede zwischen den Zentren schließen lassen (vgl. Tabelle 2.8).

Für den Fall  $H_1 : \sigma_a^2 \neq 0$  soll für verschiedene Werte von  $\sigma_a^2$  geprüft werden,

**Tabelle 2.7: Signifikanzniveau des Wald-Z-Tests für  $\sigma_a^2$**   
 ( $\sigma_a^2 = 0$ ,  $\alpha/2 = 0,025$ , Schätzmethode=REML)

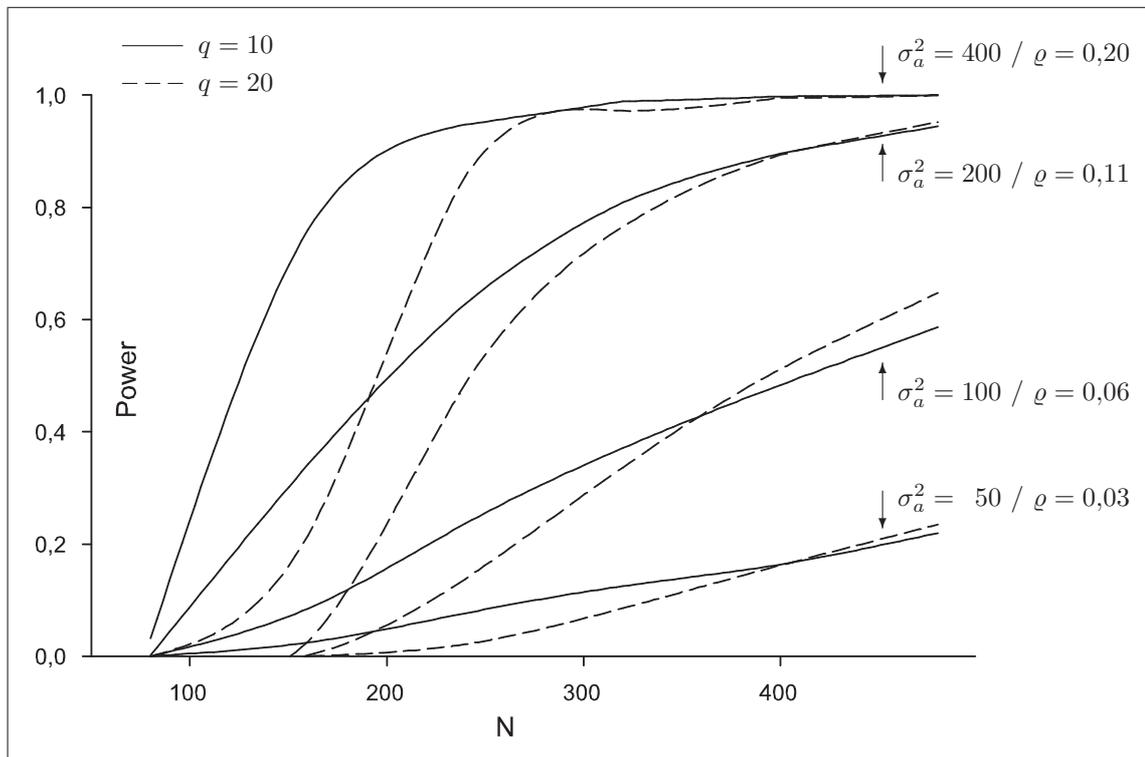
$\sigma_e^2$	$N$	$p$	balanciert		unbalanciert		
			$q$	Anteil $H_1$	$n_i$	$Q$	Anteil $H_1$
1.600	960	48	20	1,31%	10/30	19,89	1,02%
1.600	480	48	10	1,72%	5/15	9,95	1,32%
1.600	288	48	6	1,94%	3/9	5,97	1,05%
1.600	240	12	20	0,05%	10/30	19,55	0,00%
1.600	120	12	10	0,04%	5/15	9,77	0,01%
1.600	72	12	6	0,07%	3/9	5,86	0,04%
6.400	960	48	20	1,44%	10/30	19,89	0,86%
6.400	480	48	10	1,49%	5/15	9,95	1,19%
6.400	288	48	6	1,62%	3/9	5,97	1,13%
6.400	240	12	20	0,04%	10/30	19,55	0,01%
6.400	120	12	10	0,05%	5/15	9,77	0,00%
6.400	72	12	6	0,06%	3/9	5,68	0,04%

**Tabelle 2.8: Signifikanzniveau des Wald-Z-Tests für  $\sigma_a^2$ , viele Zentren**  
 ( $\sigma_a^2 = 0$ ,  $\alpha/2 = 0,025$ , Schätzmethode=REML)

$\sigma_e^2$	$N$	$p$	balanciert		unbalanciert		
			$q$	Anteil $H_1$	$n_i$	$Q$	Anteil $H_1$
1.600	960	120	8	2,75%	4/12	7,983	2,22%
1.600	960	240	4	3,90%	2/6	3,996	3,24%
1.600	960	480	2	5,15%	1/3	1,999	4,29%
6.400	960	120	8	2,96%	4/12	7,983	2,51%
6.400	960	240	4	3,64%	2/6	3,996	3,14%
6.400	960	480	2	4,61%	1/3	1,999	4,18%

wie oft die Nullhypothese zum Niveau  $\alpha/2$  verworfen wird. Die Ergebnisse für den balancierten Fall sind in Abbildung 2.7 dargestellt.

Neben der Anzahl der Zentren und der Balanciertheit des Versuchs hängt die Testgüte natürlich vom tatsächlichen Wert von  $\varrho$  und von der Gesamtfallzahl  $N$  ab. Bei jeweils gleichem  $\varrho$  ist die Testgüte bei kleineren  $N$  und bei kleinerer Wiederholungszahl  $q$  (Zentrenzahl  $p$  hoch) höher als bei höherer Zentrumsgröße ( $p$  klein).

Abbildung 2.7: Güte des Wald-Z-Tests für  $\sigma_a^2$  (balancierter Fall,  $\sigma_e^2 = 1.600$ )

Mit wachsender Gesamtfallzahl  $N$  und damit wachsender Testgüte werden die Unterschiede kleiner. Die Güte des Wald-Tests wird somit eher durch die Zahl der Zentren als durch die höhere Zahl der Patienten pro Zentrum beeinflusst. Im unbalancierten Fall sind die beobachteten Effekte – bei insgesamt etwas geringerer Güte – ähnlich (hier nicht dargestellt).

### Bemerkungen

Der verwendete Wald-Test prüft zum zweiseitigen Niveau  $\alpha$ , ob der Streuungsparameter  $\sigma_a^2$  von 0 verschieden ist, berücksichtigt bei der REML-Methode jedoch nur positive Schätzungen, wodurch das Signifikanzniveau in der Praxis nur zur Hälfte ausgeschöpft wird. Da in der Praxis der Einrichtungsvergleiche nur positive Werte von  $\sigma_a^2$  relevant sind, kann für den Test auch ein einseitiges Testniveau von beispielsweise  $\alpha = 5\%$  (entspricht einem zweiseitigen  $\alpha$  von 10%) spezifiziert werden.

Eine Methode, um zum tatsächliche Niveau- $\alpha$ -Tests darzustellen, könnten mittels Simulationen für eine gegebene Beispielsituation ( $p$  und  $n_i$  bekannt) empirische

Verteilungen der Z-Statistiken unter  $H_0 : \sigma_a^2 = 0$  bestimmt und anhand dieser kritische Grenzen ermittelt werden. Dieses Verfahren wurde für das Anwendungsbeispiel dieser Arbeit durchgeführt (vgl. Abschnitt 2.4.2.3).

### 2.3.5.5 Eigenschaften von Klinik-spezifischen Konfidenzintervallen und Hypothesentests

Ein wichtiges Ziel von Einrichtungsvergleichen ist es, Aussagen über einzelne Kliniken hinsichtlich ihrer individuellen Effekte auf den Behandlungserfolg zu treffen und zu entscheiden, ob sich dieser (signifikant) von 0 unterscheidet. Somit sind – abgesehen von Testentscheidungen über den globalen Streuungsparameter  $\sigma_a^2$  – vorwiegend Aussagen über die Punktschätzer der einzelnen Einrichtungen und deren Vertrauensbereiche von Interesse. Wie in den vorigen Abschnitten festgestellt wurde, sind die EBLUP-Punktschätzer für die Erwartungswerte der einzelnen Kliniken zwar konservativ (d.h. ihre Variation ist geringer als der wahre Streuungsparameter  $\sigma_a^2$ ), jedoch sind bisher keine Aussagen über die Lage der korrespondierenden Vertrauensbereiche und der damit verbundenen Testentscheidungen getroffen worden.

Zu diesem Zweck werden im gewählten Beispiel nun die Zentrumseffekte entsprechend des vordefinierten  $\sigma_a^2$  generiert. Dabei werden wieder die beiden Fälle  $H_0$  ( $\sigma_a^2 = 0$ , d.h. alle Zentren besitzen denselben Erwartungswert  $\mu_i = \mu$  bzw.  $\gamma_i=0$ ) und  $H_1$  ( $\sigma_a^2 > 0$ ) betrachtet.

Für den Wald-Z-Test wurde jeweils ein einseitiges Signifikanzniveau von 5% gewählt. Nur im Falle der – unter  $H_0$  fälschlichen und unter  $H_1$  korrekten – Ablehnung von  $H_0$  wurden dann zweiseitige 95%-Konfidenzintervalle für die  $\hat{\gamma}_i$  betrachtet und mit dem Wert 0 verglichen. Zur Bestimmung der Freiheitsgrade für die Konfidenzintervalle wurde die Methode nach Kenward and Roger [32] (vgl. Seite 84) verwendet. Hierbei wird für jedes Klinikum der Hypothesensatz

$$H_{0i} : \gamma_i = 0 \quad \text{vs.} \quad H_{1i} : \gamma_i \neq 0 \quad \text{mit } i = 1, \dots, p$$

zugrunde gelegt. Diese  $p$  Nullhypothesen werden getestet, indem geprüft wird, ob das  $(1-\alpha)$ -Vertrauensintervall der Form  $[k_{iu}; k_{io}]$  für  $\hat{\gamma}_i$  den Wert 0 einschließt oder nicht. Im letzteren Fall wird der Erwartungswert des Zentrums  $i$  zum zweiseitigen Niveau  $\alpha$  als verschieden von 0 eingestuft. Zur Entscheidung bezüglich eines Zentrumseffekts  $\gamma_i$  können somit die folgenden drei Fälle auftreten:

- (−1)  $k_{io} < 0$  das Konfidenzintervall liegt gänzlich unterhalb von 0  
 → Entscheidung für  $H_{1i}$  ( $\gamma_i < 0$ );
- ( 0)  $0 \in [k_{iu}; k_{io}]$  das Konfidenzintervall enthält den Wert 0  
 → Annahme von  $H_{0i}$  ( $\mu_i = \mu$ ); und
- (+1)  $k_{iu} > 0$  das Konfidenzintervall liegt gänzlich oberhalb von 0  
 → Entscheidung für  $H_{1i}$  ( $\gamma_i > 0$ ).

Für die Simulationen wurden die folgenden Fälle ausgewählt:

**Nullhypothese:** Im Fall von  $\sigma_a^2 = 0$  wird das Signifikanzniveau – wie im letzten Abschnitt festgestellt – in vielen Situationen nicht ausgeschöpft, d.h. dass der Anteil der Testentscheidungen für Unterschiede zwischen den Kliniken geringer als das vordefinierte Testniveau  $\alpha$  ist. Laut Tabelle 2.7 wird das Signifikanzniveau bei Versuchen mit hoher Zentren-Anzahl und geringerer Wiederholungszahl zumindest annähernd ausgeschöpft. Daher wird ein solcher Fall mit  $p = 48$  und  $q = 10$  ( $N = 480$ ), für die nachfolgend dargestellte Beispiel-Simulation ausgewählt.

**Alternativhypothese:** Für den Fall  $\sigma_a^2 > 0$  sei eine Kombination aus  $\sigma_a^2$  und  $\sigma_e^2$  gewählt, bei der die Testgüte des Wald-Z-Tests hoch (d.h.  $> 95\%$ ) ist. In dem balancierten Fall  $N = 480$ ,  $p = 48$  ( $q = 10$ ) und  $\varrho = 0,11$  kann dies als erfüllt angesehen werden. Hier lag die beobachtete Testgüte für  $\alpha/2 = 0,025$  bei  $\approx 95\%$  (vgl. Abbildung 2.7).

Die Ergebnisse sind in Tabelle 2.9 nach dem wahren Erwartungswert  $\mu_i = \mu + \gamma_i$  in Klassen mit einer Breite von 10 dargestellt. Für den wahren Gesamtmittelwert wird wieder  $\mu = 100$  angenommen. Die Darstellung der Häufigkeiten (basierend auf jeweils 100.000 Simulationsläufen für  $H_0$  und  $H_1$ ) erfolgte auf Basis aller Versuche und alternativ auf Basis der Fälle, in denen  $H_0$  tatsächlich abgelehnt wurde. Nur in diesen Fällen kamen die Einzelhypothesen  $H_{0i}$  zur Prüfung. Aufgrund der zu geringen absoluten Anzahl und der damit verbundenen geringen Sicherheit wurden Erwartungswerte  $\mu_i < 35$  bzw.  $\mu_i > 165$  nicht dargestellt. Der verbleibende Wertebereich enthält jedoch 99,9996% aller Fälle für diese Situation. Auch für  $\mu_i \approx 40$  bzw.  $\mu_i \approx 160$  sind die Ergebnisse recht ungenau, da hier bei 100.000 Simulationen à 48 Zentren jeweils nur etwa 230 Fälle in das jeweilige Intervall entfielen.

**Tabelle 2.9: Simulationsergebnisse zu Klinik-spezifischen Konfidenzintervallen und Hypothesentests ( $\alpha/2 \approx 0,05$ , Schätzmeth.=REML)**

Fall*	$\mu_i$	$E(h(\mu_i))$	Anteil an allen Versuchen			Anteil nach Entsch. „ $H_1$ “		
			(-1)	(0)	(+1)	(-1)	(0)	(+1)
$H_0$	100	1,0000	0,015%	5,755%	0,014%	0,253%	99,499%	0,248%
$H_1$	40 ± 5	0,0000	98,198%	1,802%	0,000%	98,198%	1,802%	0,000%
$H_1$	50 ± 5	0,0007	83,772%	15,922%	0,000%	84,029%	15,971%	0,000%
$H_1$	60 ± 5	0,0059	61,005%	38,213%	0,000%	61,486%	38,514%	0,000%
$H_1$	70 ± 5	0,0319	33,233%	65,569%	0,000%	33,636%	66,364%	0,000%
$H_1$	80 ± 5	0,1059	12,857%	85,469%	0,003%	13,076%	86,921%	0,003%
$H_1$	90 ± 5	0,2174	3,267%	94,595%	0,055%	3,337%	96,607%	0,057%
$H_1$	100 ± 5	0,2763	0,532%	96,720%	0,534%	0,544%	98,910%	0,546%
$H_1$	110 ± 5	0,2174	0,057%	94,585%	3,288%	0,058%	96,585%	3,357%
$H_1$	120 ± 5	0,1059	0,004%	85,495%	12,795%	0,004%	86,978%	13,017%
$H_1$	130 ± 5	0,0319	0,000%	65,268%	33,507%	0,000%	66,077%	33,923%
$H_1$	140 ± 5	0,0059	0,000%	38,129%	61,138%	0,000%	38,410%	61,590%
$H_1$	150 ± 5	0,0007	0,000%	15,248%	84,420%	0,000%	15,298%	84,702%
$H_1$	160 ± 5	0,0000	0,000%	3,404%	96,596%	0,000%	3,404%	96,596%

\* $H_0$ :  $\sigma_a^2 = 0$ ,  $\sigma_e^2 = 1.600$ ,  $N = 480$ ,  $p = 48$ ,  $q = 10$ , Testgüte: 5,78%

\* $H_1$ :  $\sigma_a^2 = 200$ ,  $\sigma_e^2 = 1.600$ ,  $N = 480$ ,  $p = 48$ ,  $q = 10$ , Testgüte: 98,04%

**Nullhypothese:** Für  $\sigma_a^2 = 0$  (siehe erste Datenzeile Tabelle 2.9) sind in 5,78% der 100.000 Beispielstudien Unterschiede zwischen den Kliniken festgestellt worden, wenn keine Unterschiede vorliegen (das vorgegebene Signifikanzniveau von 5% wurde somit leicht überschritten). In 0,253% dieser Fälle wurde ein Klinik-Effekt  $\gamma_i$  fälschlicherweise als signifikant  $< 0$  eingestuft. Die Gesamtwahrscheinlichkeit – d.h. auf Basis aller Versuche – für eine Klinik, fälschlicherweise als im Effekt verschieden von 0 klassifiziert zu werden, beträgt approximativ 0,03%. Der Fall, dass zwar Unterschiede zwischen den Kliniken gefunden werden, aber kein Klinik-Effekt als signifikant verschieden von 0 eingestuft wird, ist hier – unabhängig von  $\sigma_e^2$  – durchaus häufig (und gilt natürlich auch für echt positive  $\sigma_a^2$ ).

**Alternativhypothese:** Bei Vorliegen von  $\sigma_a^2 = 200$  wurden in etwas mehr als 98% aller 100.000 simulierten Studien durch Betrachtung von  $\hat{\sigma}_a^2$  Unterschiede zwischen den Zentren festgestellt. Einrichtungen mit einem wahren Erwartungswert von beispielsweise etwa 80 (d.h.  $\gamma_i \in [-25; -15[$  bzw.  $\mu_i \in [75; 85[$ ) wurde

in etwa 12,9% aller Versuche korrekterweise auf  $H_{1i}$  ( $\gamma_i < 0$ ) geschlossen. Bei ca. 85,5% derjenigen Fälle, in denen ein solcher Klinikeffekt vorhanden war, wurde dieser Unterschied nicht aufgedeckt. In 0,002% aller Versuche wurde für ein solches Zentrum fälschlicherweise auf  $H_{1i}$  ( $\gamma_i > 0$ ) geschlossen. Eine hohe Testgüte für  $H_{0i}$  von mehr als 80% ist in der gewählten Situation erst bei recht hohen Abweichungen (d.h.  $|\gamma_i| > 45$ ) gegeben.

**Bemerkung:**

Die Trennschärfe der Klinik-spezifischen Tests für die  $H_{0i}$  hängt neben der Restfehlerstreuung  $\sigma_e^2$  (vgl. vorausgehender Absatz) natürlich in hohem Maße von der Wiederholungszahl  $q$  bzw.  $n_i$  ab. Zur Illustration des Effekts sei exemplarisch die gleiche Situation, wie sie für die Tabelle 2.9 gewählt wurde, betrachtet, jedoch mit  $q = 20$  anstelle von  $q = 10$ . In diesem Falle beträgt die Güte des Globatests bereits  $\approx 100\%$  (1.000 Simulationsläufe). Die Testgüte für ein Zentrum mit einem Erwartungswert von etwa 80 liegt hier bei approximativ 39%.

**Schlussbemerkung**

Für die in diesem Abschnitt dargestellten Zentrumseffekte  $\gamma_i$  gilt zwar

$$\hat{\mu}_i^{\text{EBLUP}} = X\hat{\beta} + Z\hat{\gamma} = \hat{\mu} + \hat{\gamma}_i^{\text{EBLUP}},$$

jedoch können die Konfidenzgrenzen für die  $\hat{\gamma}_i$  nicht für die  $\hat{\mu}_i$  verwendet werden. Zur Konstruktion der Konfidenzintervalle (mittels t-Verteilung) für die  $\hat{\gamma}_i$  werden die Prediktionsfehler  $\hat{\gamma}_i - \gamma_i$ , die gerade durch die EBLUP-Schätzung gemeinsam minimiert werden, verwendet. Zur Konstruktion von Konfidenzgrenzen für die  $\hat{\mu}_i$  muss zusätzlich die Variation von  $\hat{\beta}$ , die von Stichprobenumfang und der Restfehlerstreuung abhängt, berücksichtigt werden.

## 2.4 Generalisierte Gemischte Lineare Modelle

### 2.4.1 Einführung

Wie anhand des Beispiels der logistischen Regression eingeführt, können Zielgrößen mit nicht-stetigem Skalenniveau mittels einer Link-Funktion  $g$  in analoger Weise zum Linearen Modell dargestellt werden. Im Falle fester und zufälliger Effekte schreibt sich das Modell dann zu

$$\begin{aligned} g(y) &= X\beta + Z\gamma && \text{bzw.} \\ E(y|Z) &= g^{-1}(X\beta + Z\gamma) . \end{aligned}$$

Ein Modell dieser Form heißt „Generalisiertes Gemischtes Lineares Modell“ („Generalized Linear Mixed Model“ [GLMM]). Der Term  $X\beta + Z\gamma$  heißt in diesem Zusammenhang auch **LINEARER PREDIKTOR**.

Die bei Einrichtungsvergleichen vorliegende Klumpenstruktur kann wieder mittels eines hierarchischen („multilevel“) Modells angepasst werden (vgl. Fahrmeir und Tutz [12] oder Verbeke und Molenberghs [57]), indem Faktoren mit festen Effekten (wie beispielsweise demographische und anamnestiche Kovariaten der Patienten, siehe auch Kapitel 3) über das Matrizenprodukt  $X\beta$  und die Einflussgröße *Einrichtung* als Faktor mit zufälligen Effekten in  $Z\gamma$  modelliert werden. Hierbei werden die Subjekte innerhalb der Einrichtungen – bei gegebenem Klinik-spezifischen Erwartungswert (bzw. eine Ereigniswahrscheinlichkeit) – als voneinander unabhängige Beobachtungen der ersten Stufe („level-1“) betrachtet.

Ein solches Modell mit einem Faktor, der zufällige Effekte berücksichtigt, hat bei binärer Zielgröße die folgende Gestalt:

$$g(y_{ik}) = \text{logit}(p_{ij}) = \left( \beta_0 + \sum_{j=1}^k \beta_j X_{ij} \right) + \gamma_i + e_{ik} ,$$

mit  $\gamma_i$  als Koeffizienten für die  $i$ -te Einrichtung. Für die Ereigniswahrscheinlichkeiten in einem Klumpen (Klinikum)  $i$  gilt folglich

$$E(y_i|\gamma_i) = P(Y = 1|\gamma_i) = g^{-1}(X_i\beta + \gamma_i) .$$

Die nächst höhere (zweite) Stufe bilden somit die Einheiten, innerhalb derer die Subjekte (hier Patienten) betrachtet werden. Diese können auch als Messwiederholungen innerhalb ein und derselben Ausprägung der zweiten Stufe, hier der Einrichtung, interpretiert werden. Die Abweichungen der Einzelmessungen von ihrem Erwartungswert  $e_{ik}$  nennt man auch „Level-1 Residuen“.

Für die zufälligen Effekte auf der zweiten Stufe wird nun wieder Unabhängigkeit mit Erwartungswert 0 vorausgesetzt. Häufig wird weiterhin angenommen, dass die Effekte einer Normalverteilung mit Varianz  $\sigma_a^2$  entstammen. Die Abweichungen der  $\gamma_i$  von ihrem Erwartungswert 0 nennt man entsprechend „Level-2 Residuen“.

Generalisierte Gemischte Lineare Modelle können darüber hinaus auch in der Variante des marginalen Modells dargestellt werden. Hier werden die Erwartungswerte  $\mu_i$  bzw.  $p_i$  *unbedingt* modelliert,

$$E(y_i) = g^{-1}(X_i\beta) ,$$

jedoch mit einer Varianzfunktion  $V(y_i)$  versehen. Die Kovarianz der Paare  $(y_i, y_j)$  wird über eine Funktion  $C(\mu_i, \mu_j; \alpha)$  geschätzt. In diesem Ansatz werden somit eher populationsbezogene Zusammenhänge betrachtet, welche im Rahmen von Einrichtungsvergleichen weniger von Interesse sind. Daher wird im Rahmen dieser Arbeit auf das hierarchische Modell Bezug genommen.

Zur Schätzung der festen und zufälligen Komponenten in generalisierten linearen Modellen stehen verschiedene auf restringierten Pseudo-Likelihood-Verfahren (RPL) basierende Methoden zur Verfügung:

- RSPL: Residual Subject-specific Pseudo-Likelihood;
- MSPL: Maximum Subject-specific Pseudo-Likelihood;
- RMPL: Residual Marginal Pseudo-Likelihood;
- MMPL: Maximum Marginal Pseudo-Likelihood.

Diese Schätzmethoden unterscheiden sich im Wesentlichen durch zwei Faktoren:

- die Stelle der Expansion bei der Linearisierung (population averaged (PA) oder subject specific (SS)),
- die Zielfunktion (residual oder maximum pseudo-likelihood).

Die RSPL-Methode beispielsweise beruht somit auf dem hierarchischen Modell (Expansionstelle  $\tilde{\gamma} = \hat{\gamma}$ ), besitzt für die Schätzung von  $\gamma$  in den meisten Situationen die besten Eigenschaften und wird häufig als Standardmethode im GLMM vorgeschlagen.

Zur Optimierung der Zielfunktionen wurden in der Literatur wiederum vielfältige Algorithmen vorgeschlagen, von denen hier exemplarisch die „Quasi-Newton Optimierung“ für die RSPL-Schätzung genannt sei. Bei dieser Methode wird – analog zum linearen gemischten Modell – eine Untergrenze von 0 für die Variationskomponenten vorgegeben. Weitere Details zu Schätzung und Optimierung im GLMM sind bei Wolfinger und O’Connell [61] nachzulesen.

### Anpassungsgüte im GLMM

Obwohl Modell-Optimierung nicht im Mittelpunkt dieser Arbeit steht, sei abschließend ein Kriterium genannt, mittels dessen die Anpassungsgüte generalisierter gemischter linearer Modelle gemessen werden kann. Der Term

$$\chi_g^2 = \hat{r}' V(\hat{\theta})^{-1} \hat{r} \quad (2.15)$$

heißt „generalisierte Chi-Quadrat-Statistik“. Zur Terminologie sei auf das einführende Kapitel 2.3.1 bzw. die Gleichungen (2.2) und (2.3) verwiesen.

Das Verhältnis aus  $\chi_g^2$  und seiner Zahl der Freiheitsgrade gibt die Variabilität der Residuen im marginalen Modell an.

## 2.4.2 Streuungsursachen bei generalisierten gemischten linearen Modellen

### 2.4.2.1 Motivation, Auswirkung zufälliger Effekte

Ähnlich wie bei stetig normalverteilten Merkmalen existiert bei binären Zielgrößen in Situationen mit zufälligen Effekten (bei hierarchischer Datenstruktur) eine Varianzaufblähung der beobachteten nach den Zentren aggregierten Ergebnissen. Zur Darstellung des Effekts wird zunächst die Zielgröße  $Y_i$  als Anzahl der Ereignisse im Zentrum  $i$  modelliert, also als Summe der  $n_i$  Bernoulli-Experimente  $y_{ij}$  und somit als binomialverteilte Zufallsvariable aufgefasst mit Wahrscheinlichkeitsparameter  $p_i$  und  $n_i$  Versuchen. Hierbei gilt  $p_i$  in einem einfachen logistischen Regressionsmodell innerhalb eines Zentrums für *alle* Subjekte, und somit gilt:

$$E(Y_i) = n_i p_i \quad \text{und} \quad \text{Var}(Y_i) = n_i p_i (1 - p_i) .$$

Der unbekannte Wahrscheinlichkeitsparameter kann mittels  $\hat{p}_i = y_i/n_i$  erwartungstreu geschätzt werden. In einem Einrichtungsvergleich werden gewissermaßen  $p$  Binomialverteilungen

$$B_1(n_1, p_1), \dots, B_p(n_p, p_p)$$

betrachtet. Als Realisationen der beobachteten Ereignishäufigkeiten ergibt sich zunächst eine (von der unbekanntem Verteilung der Erwartungswerte  $E(Y_1), \dots, E(Y_p)$  gänzlich verschiedene) *Mischverteilung* aus den  $p$  Wahrscheinlichkeitsverteilungen der Form

$$B(Y) = r_1 B_1(Y_1) + \dots + r_p B_p(Y_p) ,$$

wobei für die Wahl der Gewichte  $r_i$  die Fallzahlanteile  $r_i = n_i/n$  geeignet sind. Wenn  $\sum_{i=1}^p r_i = 1$ , so ist  $B(y)$  wieder eine Wahrscheinlichkeitsverteilung.

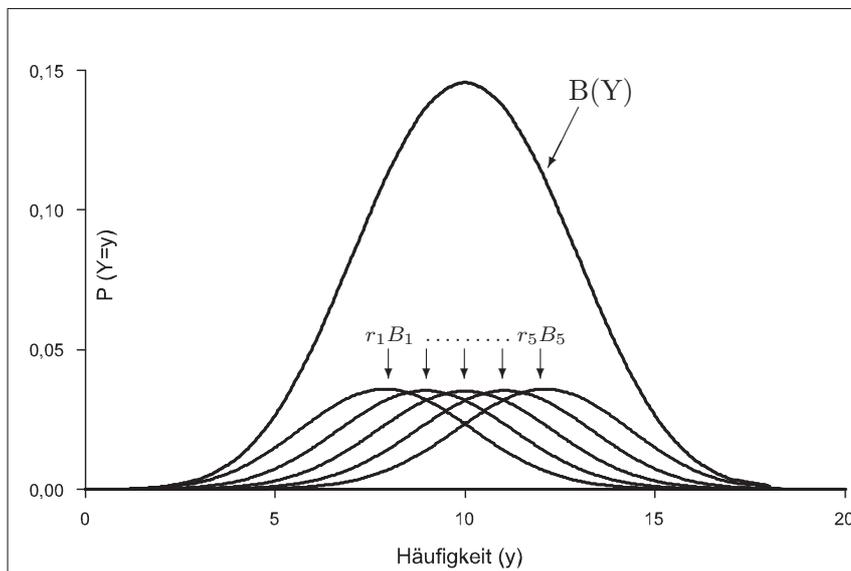
Als illustrierendes Beispiel hierfür sein ein Versuch mit  $p = 5$  Einrichtungen und balancierten Fallzahlen von jeweils  $n_i = 20 \forall i$  gewählt. Die Ereigniswahrscheinlichkeiten seien als

$$\{p_1 ; p_2 ; p_3 ; p_4 ; p_5\} = \{0,40 ; 0,45 ; 0,50 ; 0,55 ; 0,60\}$$

gegeben. Somit gilt für jedes  $k \in \{0, 1, 2, \dots, 20\}$  (vgl. Abbildung 2.8):

$$\begin{aligned} P(Y = k) &= \sum_{i=1}^p \frac{n_i}{n} P(Y_i = k) \\ &= \sum_{i=1}^5 \frac{20}{100} \binom{n_i}{k} p_i^k (1 - p_i)^{n_i - k} . \end{aligned}$$

**Abbildung 2.8: Mischverteilung von  $p = 5$  binomialverteilten Zufallsgrößen**



Die Varianzen der einzelnen Binomialverteilungen

$$\begin{aligned} \text{Var}(B_i) &= E(Y_i - E(Y_i))^2 \\ &= \sum_{k=0}^{n_i} P(Y_i = k)(k - E(Y_i))^2 \\ &= \sum_{k=0}^{n_i} P(Y_i = k)(k - n_i p_i)^2 = n_i p_i (1 - p_i) \\ &= (4,80 ; 4,95 ; 5,00 ; 4,95 ; 4,80) \end{aligned}$$

sind geringer als die der Mischverteilung  $B(Y)$ :

$$\begin{aligned}
\text{Var}(B) &= E(Y - E(Y))^2 \\
&= \sum_{k=0}^{\max_i(n_i)} P(Y = k)(k - E(Y))^2 \\
&= \sum_{k=0}^{\max_i(n_i)} \left( \sum_{i=0}^p r_i P(Y_i = k) - E(Y) \right)^2 \\
&= \sum_{k=0}^{\max_i(n_i)} \left( \sum_{i=0}^p r_i P(Y_i = k) - \frac{1}{n} \sum_{i=1}^p r_i n_i p_i \right)^2 \\
&= 6,90 .
\end{aligned}$$

Somit ist natürlich auch die Varianz der Erwartungswerte  $\mu_i = n_i p_i$  (im Beispiel = 2,5) geringer als die erwartete Varianz der beobachteten  $Y_i$ . Dieses Phänomen ist in Analogie zum normalverteilten Fall zu sehen (Varianz-Exzess), wo sich ähnliche Mischverteilungen aus den unterschiedlichen Messreihen ergeben.

Die Varianz des Stichprobenmittels  $\hat{\mu} = \bar{y}$ , welches natürlich selbst eine Zufallsvariable ist, beträgt im balancierten Fall

$$\begin{aligned}
\text{Var}(\hat{\mu}) &= \text{Var} \left( \frac{1}{p} (Y_1 + Y_2 + \dots + Y_p) \right) \\
&= \frac{1}{p^2} (\text{Var}(Y_1) + \text{Var}(Y_2) + \dots + \text{Var}(Y_p)) \\
&= 0,98 .
\end{aligned}$$

Im unbalancierten Fall ergibt sich die Varianz des Stichprobenmittels  $\hat{\mu} = \sum_{i=1}^p r_i Y_i$  mit den Gewichten  $r_i = n_i/n$  zu

$$\text{Var}(\hat{\mu}) = \text{Var} \left( \sum_{i=1}^p r_i Y_i \right) = \sum_{i=1}^p r_i^2 \text{Var}(Y_i)$$

und verhält sich somit umgekehrt proportional zur Anzahl der Einrichtungen und proportional zur (durchschnittlichen) Fallzahl. Diese Variabilität wird durch die Varianz zwischen den beobachteten  $y_i$  der Subpopulationen verursacht und beträgt (vgl. Nebenrechnung im Anhang B.1)

$$\begin{aligned}
\text{Var}(y_1, \dots, y_p) &= E \left( \frac{1}{p-1} \sum_{i=1}^p (y_i - \bar{y})^2 \right) \\
&= p \text{Var}(\hat{\mu}) + \text{Var}(E_1, \dots, E_p) \text{ (balancierter Fall), bzw.} \\
&= \frac{1}{p} \sum_{i=1}^p \text{Var}(Y_i) + \text{Var}(E_1, \dots, E_p) \text{ (unbalancierten Fall).}
\end{aligned}$$

Diese Varianz ist in der balancierten Situation wiederum um die Konstante

$$\frac{\text{Var}(E_i)}{p}$$

größer ist als die der Verteilung  $B$ ,

$$\text{Var}(y_1, \dots, y_p) = \text{Var}(B) + \frac{\text{Var}(E_1, E_2, \dots, E_p)}{p},$$

und strebt somit mit wachsender Anzahl der Zentren bei konstanter Variabilität zwischen den Zentren gegen die Variabilität der Mischverteilung:

$$\text{Var}(y_1, \dots, y_p) - \text{Var}(B) \xrightarrow{p \rightarrow \infty} 0.$$

Beim Vergleich zwischen Einrichtungen sind nicht primär die absoluten Ereignishäufigkeiten, sondern eher die Anteile ( $h_i = y_i/n_i$ ) von Interesse. Die Variation des (ungewichteten) Stichprobenanteils ( $\hat{p} = \bar{h}_i$ ) ist unter Verwendung der Einzelvarianzen ( $\text{Var}(h_i) = p_i(1-p_i)/n_i$ )

$$\begin{aligned}
\text{Var}(\hat{p}) &= \text{Var} \left( \frac{1}{p} (h_1 + h_2 + \dots + h_p) \right) \\
&= \frac{1}{p^2} (\text{Var}(h_1) + \text{Var}(h_2) + \dots + \text{Var}(h_p)) \\
&= 0,0025,
\end{aligned}$$

und im unbalancierten Fall für  $\hat{p} = \sum_{i=1}^p r_i h_i$

$$\text{Var}(\hat{p}) = \text{Var}\left(\sum_{i=1}^p r_i h_i\right) = \sum_{i=1}^p r_i^2 \text{Var}(h_i).$$

Für die Variation der relativen Ereignishäufigkeiten  $h_i$  gilt allgemein

$$\text{Var}(h_1, \dots, h_p) = p \text{Var}(\hat{p}) + \text{Var}(p_1, \dots, p_p)$$

und für das oben beschriebene Beispiel

$$\text{Var}(h_1, \dots, h_5) = 5 * 0,00245 + 0,00611 = 0,0184.$$

### Varianzdispersion

Zusätzlich zur im vorigen Abschnitt beschriebenen Varianzaufblähung existiert bei binären Daten ein Effekt, der dazu führt, dass die Realisationen  $y_i$  sogar eine noch größere Varianz aufweisen können, als durch die (theoretische) Mischverteilung  $B(Y)$  gegeben. Diesen Effekt nennt man „Überdispersion“ („overdispersion“) oder „Extra-Binomial-Variation“. Dieser Effekt kann durch Störeinflüsse entstehen, die nicht vom Modell abgedeckt werden (Fehl- oder Unterspezifizierung des Modells). Diese Einflüsse führen dazu, dass der Wahrscheinlichkeitsparameter  $p_i$  nicht für alle Subjekte innerhalb des  $i$ -ten Zentrums gleichermaßen, sondern nur im Mittel gilt (siehe auch McCullagh und Nelder [39]). Im Gegensatz zum normalverteilten Fall kann dieser Effekt jedoch durch das Modell nicht direkt geschätzt werden, da es hier keinen vom zu schätzenden Parameter  $p_i$  selbst unabhängigen Streuungsparameter gibt (vgl. Kapitel 2.2.4).

Diese zusätzliche Streuungskomponente kann jedoch durch einen *multiplikativen* so genannten „Scale“- oder Dispersionsparameter  $\phi$  dargestellt und geschätzt werden, um den die beobachteten absoluten Ereignishäufigkeiten  $y_i$  stärker variieren werden als durch das Modell unterstellt:

$$\text{Var}(Y_i) = \phi n_i \tilde{p} (1 - \tilde{p}).$$

Das Ausmaß dieses Dispersionsfaktors hängt somit von der Klumpengröße und der Variabilität der wahren  $p_i$  ab und kann bei ausreichender Klumpengröße und -anzahl durch Verwendung der Pearson'schen  $\chi^2$ -Statistik

$$\hat{\phi} = \frac{\chi^2}{N - p}$$

geschätzt werden. Bei entsprechendem Ausmaß kann nun die Kovarianzmatrix der Modellparameter mit dem Faktor  $\hat{\phi}$  versehen werden, um diesen Effekt auszugleichen. Ein alternativer Ansatz wird von Browne et al. [5] beschrieben, wo ein *additiver* Dispersionsparameter verwendet wird, der als Varianzpartitionierung – wie beim normalverteilten Fall – verstanden wird. Dieser hat den Vorteil, dass die konventionellen Likelihood-Methoden angewandt werden können.

Als Alternative zur nachträglichen Streuungsadjustierung kann aber ein gemischtes Modell angepasst werden, wodurch die Schätzung weiterer Streuungskomponenten (wie etwa die der Klumpenvariablen selbst) mit zufälligen Effekten möglich wird. In jüngster Zeit sind entsprechende Schätzverfahren in den gängigen Software-Paketen verfügbar geworden (vgl. nachfolgender Abschnitt und Kapitel 3.6).

#### 2.4.2.2 Empirische Eigenschaften von Modellschätzern und Tests bei binomialverteilter Zielgröße (Logit-Modell)

Im Kapitel 2.3.5.3 wurden für die normalverteilte Situation gewisse empirische Eigenschaften von EBLUP-Schätzern und Varianzkomponenten-Tests mittels Simulationen geprüft. Entsprechendes soll nun in verkürzter Form für den binären Fall dargestellt werden, da die Zielgröße des Beispieldatensatzes von dichotomem Skalenniveau ist.

Das hier zugrunde gelegte generalisierte gemischte lineare Modell mit gewählter Logit-Linkfunktion hat somit die folgende Gestalt (mit  $p := P(Y = 1)$ ):

$$g(p) = \text{logit}(p) = \ln\left(\frac{p}{1-p}\right) = X\beta + Z\gamma + e = \beta_0 + \gamma_i + e.$$

Die linearen Prediktoren für die Zentrumswahrscheinlichkeiten im Logit-Modell lauten wieder

$$\hat{\mu}_i^{\text{EBLUP}} = X\hat{\beta} + Z\hat{\gamma} = \hat{\beta}_0 + \hat{\gamma}_i$$

und  $\hat{p}_i = \hat{P}(Y = 1|\gamma_i) = \frac{1}{1 + \exp(-\hat{\mu}_i)}.$

Die Eigenschaften der Modell- und EBLUP-Schätzungen für die Ereigniswahrscheinlichkeiten werden nun anhand von Simulationen in der einfachen hierarchischen Klassifikation im Logit-Modell betrachtet. Hierzu wurde jetzt eine Datensituation gewählt, die dem Anwendungsbeispiel entspricht (vgl. Kapitel 4). Es werden daher  $p = 10$  Einrichtungen mit den folgenden Fallzahlen gewählt (vgl. Tabelle 4.3):

$$\begin{aligned} \{n_1, \dots, n_{10}\} &= \{342, 639, 454, 223, 560, 209, 200, 338, 267, 233\} \\ \implies N &= 3.465 . \end{aligned}$$

Die Kliniknummern aus der Datenbank wurden zum Zwecke der Verblindung vermischt und von 1 bis 10 neu vergeben.

Die Zentrumseffekte  $\gamma_i$  im linearen Prediktor werden wieder als normalverteilt angenommen ( $\gamma_i \sim N(0, \sigma_a^2)$ ). Für die Gesamtpopulation gelte die im Anwendungsbeispiel beobachtete Grundwahrscheinlichkeit für das Ereignis ( $Y = 1$ ):

$$\tilde{p} = \frac{397}{3.465} = 0,114574 \quad \implies \quad \beta_0 = -2,0448 \text{ im Logit-Modell.}$$

Da im binomialen Fall des GLMM nur ein Streuungsparameter – nämlich  $\sigma_a^2$  – existiert, existiert hier kein Intra-Klassen-Korrelationskoeffizient  $\rho$  und keine Wiederholbarkeits-Quotienten  $\rho_i$ .

Für die Variation der Zentrumseffekte wurden die in der Tabelle 2.11 dargestellten Situationen für den Streuungsparameter  $\sigma_a^2$  im linearen Prediktor gewählt. Die resultierenden Variationen der Ereignishäufigkeiten und -wahrscheinlichkeiten wurden mit jeweils 100.000 Simulationen mit der mit Stand von Juni 2006 neu entwickelten Software PROC GLIMMIX („Generalized Linear Mixed Models“, Fa. SAS<sup>®</sup> Institute) Schätzungen für  $\sigma_a^2$ ,  $X\hat{\beta} = \hat{\beta}_0$  und  $Z\hat{\gamma} = \hat{a}_i$  ( $i = 1, \dots, 10$ ) mittels RSPL-Methoden bestimmt. Für die Simulation wurden die Konvergenz-Kriterien für die Zielfunktion manuell schwächer als die Standardeinstellung eingestellt (mittels NLOPTIONS ABSGCONV=0.002 1), um Programmabbrüche aufgrund nicht konvergierender Algorithmen weitestgehend auszuschließen. Mit den gewählten Kriterien (d.h. eine wenigstens einmalige Änderung der Zielfunktion gegenüber dem vorausgegangenen Iterationsschritt um absolut maximal 0,002 wurde als Konvergenz akzeptiert) konnten in jeweils mehr als 99,9% der 100.000 Simulationsläufe Modellschätzer bestimmt werden.

Zur Simulation wurden stets die folgenden Methoden und Einstellungen verwendet:

Linkfunktion: Logit,  
 Schätzmethode: Residual Subject-specific Pseudo-Likelihood (RSPL),  
 Freiheitsgrade: Containment (Residual);  $df = N - \text{Rg}(XZ)$ ,  
 Optimierungsmethode: Quasi-Newton;  
 Konvergenzkriterium: 0,002 (absolut).

**Tabelle 2.10: Simulationsergebnisse zu Ereignishäufigkeiten im Random-Logit-Modell, unbalanciert ( $N = 3.465$ ,  $p = 10$ ,  $\tilde{p} = 0,1146$ )**

$\sigma_a^2$	$h = \sum y_{ij}/N^*$	beobachteter Mittelwert für					
		$Var(p_i)$	$p_{(1)}$	$p_{(10)}$	$Var(h_i)$	$h_{(1)}$	$h_{(10)}$
0,00 ( $H_0$ )	0,1146	0	0,1146		0,0003	0,0866	0,1441
0,01 ( $H_1$ )	0,1150	0,0001	0,1000	0,1313	0,0004	0,0836	0,1489
0,02 ( $H_1$ )	0,1153	0,0002	0,0945	0,1388	0,0006	0,0809	0,1534
0,03 ( $H_1$ )	0,1157	0,0003	0,0905	0,1449	0,0007	0,0785	0,1576
0,04 ( $H_1$ )	0,1162	0,0004	0,0873	0,1503	0,0008	0,0766	0,1617
0,05 ( $H_1$ )	0,1165	0,0005	0,0845	0,1550	0,0009	0,0747	0,1655
0,06 ( $H_1$ )	0,1169	0,0006	0,0821	0,1596	0,0010	0,0729	0,1629
0,08 ( $H_1$ )	0,1177	0,0009	0,0781	0,1629	0,0012	0,0699	0,1765
0,10 ( $H_1$ )	0,1185	0,0011	0,0747	0,1755	0,0014	0,0672	0,1833
0,12 ( $H_1$ )	0,1192	0,0013	0,0718	0,1826	0,0017	0,0649	0,1898
0,15 ( $H_1$ )	0,1204	0,0017	0,0679	0,1927	0,0020	0,0617	0,1991
0,20 ( $H_1$ )	0,1222	0,0023	0,0628	0,2079	0,0027	0,0573	0,2136

$$* h = \sum y_{ij} / N \approx \sum h_i / p$$

Da die Verteilung der Zentrumseffekte nur im linearen Prediktor, also im Argument der Logit-Linkfunktion  $g(\cdot)$ , symmetrisch ist, ergeben sich beim Übergang zum Skalenniveau der Zielgröße (hier Wahrscheinlichkeiten) im allgemeinen (d.h. falls  $\tilde{p} \neq 0,5$ ) asymmetrische Verteilungen. Diese sind beispielhaft in Tabelle 2.10 dargestellt. Mit steigender Variation ändert sich – im Gegensatz zum LMM – somit auch die mittlere beobachtete relative Gesamthäufigkeit des Ereignisses bei gegebenem  $\tilde{p}$ . Ist  $\tilde{p} < 0,5$ , so steigt diese; im Falle dass  $\tilde{p} > 0,5$ , sinkt sie mit wachsendem  $\sigma_a^2$ . Diese Änderung wird durch die Modellschätzung jedoch berücksichtigt und annähernd ausgeglichen (vgl. Bemerkung am Ende dieses Abschnitts).

**Tabelle 2.11: Simulationsergebnisse zu Modell-Schätzern im Random-Logit-Modell (RSPL), unbalanciert ( $N = 3.465$ ,  $p = 10$ ,  $\tilde{p} = 0,1146$ )**

$\sigma_a^2$	Mittelwert für		Schätzmittel für			Schätzvarianz für		
	$Var(\hat{\gamma}_i)$	$Var(\hat{p}_i)$	$\hat{\beta}_0$	$\hat{p}$	$\hat{\sigma}_a^2$	$\hat{\beta}_0$	$\hat{p}$	$\hat{\sigma}_a^2$
0,00 ( $H_0$ )	0,0020	0,0000	-2,0465	0,1145	0,0052	0,0029	0,0000	0,0001
0,01 ( $H_1$ )	0,0060	0,0001	-2,0443	0,1148	0,0127	0,0040	0,0000	0,0003
0,02 ( $H_1$ )	0,0115	0,0001	-2,0435	0,1149	0,0215	0,0051	0,0001	0,0005
0,03 ( $H_1$ )	0,0182	0,0002	-2,0422	0,1151	0,0309	0,0062	0,0001	0,0008
0,04 ( $H_1$ )	0,0255	0,0003	-2,0410	0,1153	0,0404	0,0072	0,0001	0,0011
0,05 ( $H_1$ )	0,0333	0,0004	-2,0404	0,1154	0,0501	0,0082	0,0001	0,0015
0,06 ( $H_1$ )	0,0418	0,0005	-2,0405	0,1154	0,0601	0,0093	0,0001	0,0019
0,08 ( $H_1$ )	0,0589	0,0007	-2,0386	0,1157	0,0797	0,0113	0,0001	0,0028
0,10 ( $H_1$ )	0,0768	0,0009	-2,0381	0,1158	0,0994	0,0133	0,0001	0,0039
0,12 ( $H_1$ )	0,0950	0,0011	-2,0377	0,1159	0,1191	0,0152	0,0002	0,0053
0,15 ( $H_1$ )	0,1235	0,0014	-2,0370	0,1161	0,1493	0,0182	0,0002	0,0076
0,20 ( $H_1$ )	0,1704	0,0020	-2,0374	0,1162	0,1983	0,0232	0,0002	0,0124

Die Varianzaufblähung zeigt sich beispielsweise im Fall von  $\sigma_a^2 = 0,05$  empirisch für die Variation der Ereigniswahrscheinlichkeiten und der beobachteten relativen Ereignishäufigkeiten (vgl. auch Kapitel 2.4.2.1):

$$Var(p_1, \dots, p_{10}) = Var\left(\frac{1}{1 + e^{-(X\beta + Z\gamma)}}\right) = Var\left(\frac{1}{1 + e^{-(\beta_0 + \gamma_i)}}\right) \approx 0,0005$$

bzw.  $Var(h_1, \dots, h_{10}) \approx 0,0009$ .

Die mittlere Spannweite der Ereigniswahrscheinlichkeiten bzw. beobachteten relativen Ereignishäufigkeiten (die auch mittels komplexerer Integrationsrechnungen über die Ordnungsstatistiken bestimmbar ist) beträgt in diesem Fall etwa

$$E(p_{(10)} - p_{(1)}) = E(p_{(10)}) - E(p_{(1)}) \approx 0,1552 - 0,0846 = 0,0708 ,$$

$$E(h_{(10)} - h_{(1)}) = E(h_{(10)}) - E(h_{(1)}) \approx 0,1660 - 0,0748 = 0,0912 .$$

Die beobachtete Schrumpfungswirkung der EBLUP-Schätzungen für die  $\gamma_i$  gegenüber den beobachteten Ereignishäufigkeiten zeigt sich durch Betrachtung der

ersten Datenspalte in Tabelle 2.11. Im Fall von  $\sigma_a^2 = 0,05$  beispielsweise beträgt die durchschnittliche Variabilität der geschätzten Effekte  $\hat{\gamma}_i$  etwa zwei Drittel der bekannten Effekt-Variabilität  $\sigma_a^2$ . Mit wachsendem  $\sigma_a^2$  erhöht sich dieser Anteil bis auf etwa 85% für  $\sigma_a^2 = 0,2$ . Dieser Umstand kann auf die mit steigendem  $\sigma_a^2$  selteneren Schätzungen mit  $\hat{\sigma}_a^2 = 0$  und den damit verbundenen Situationen mit  $\text{Var}(\hat{a}_i) = 0$  zurückgeführt werden. Damit ist ebenfalls erklärt, dass für kleine  $\sigma_a^2$  eine größere mittlere Überschätzung des Streuungsparameters vorliegt als für größere Werte von  $\sigma_a^2$ .

### Bemerkung

Die Schätzung des marginalen Modells  $\hat{\mu}_m = X\hat{\beta}$ , d.h. unter Außerachtlassung der zufälligen Effekte, nennt man bei gemischten Modellen häufig „Population Averaged (PA) Prediction“. Im Falle nicht-linearer Link-Funktionen ist  $g^{-1}(X'\beta)$  im allgemeinen nicht gleich  $E(Y)$  (wie durch die Ergebnisse in Tabelle 2.11 illustriert wurde). Mit einem Zentrums-spezifischen  $p_i = g^{-1}(X_i\beta + Z\gamma_i)$  wird somit durch  $g^{-1}(X_i\hat{\beta})$  die Ereigniswahrscheinlichkeit einer durchschnittlichen Einrichtung (d.h.  $\gamma_i = 0$ ), und nicht die Grundwahrscheinlichkeit der gesamten Population geschätzt. Bei *linearen* gemischten Modellen (bzw. Identitäts-Linkfunktion) oder auch falls  $Z\gamma = 0$  sind Mittelwert einer Durchschnitts-Einrichtung und Gesamtmittelwert identisch.

#### 2.4.2.3 Darstellung eines Signifikanztests und Betrachtungen zur Testgüte

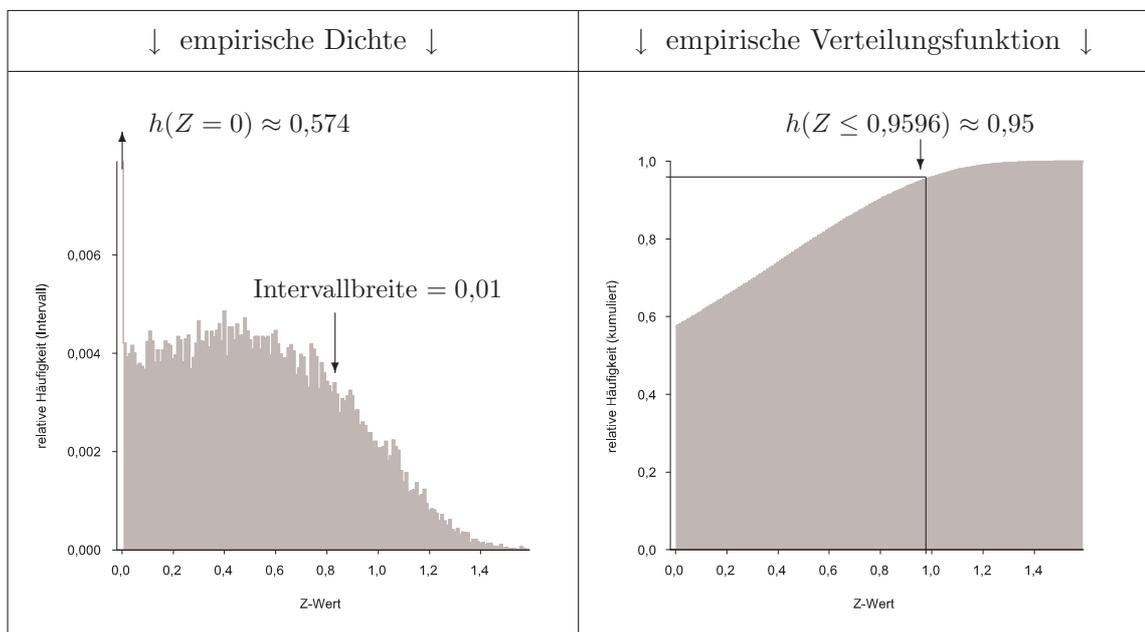
In diesem Abschnitt sollen nun wieder Eigenschaften von Signifikanztests für den Streuungsparameter  $\sigma_a^2$  diskutiert werden. Hierzu wird zunächst wieder der Fall  $H_0 : \sigma_a^2 = 0$  betrachtet. In diesem Fall sollte die globale Nullhypothese  $H_0 : \sigma_a^2 = 0$  von einem Niveau- $\alpha$ -Test in höchstens  $\alpha * 100\%$  der Versuche abgelehnt werden. In dem verwendeten Software-Paket SAS<sup>®</sup> PROC GLIMMIX werden in der aktuellen Version (Stand März 2007) zwar Schätzer und Standardfehler für den Streuungsparameter  $\hat{\sigma}_a^2$  angegeben, jedoch steht derzeit noch kein Signifikanztest für  $H_0$  zur Verfügung.

Um einen solchen Test für die gewählte Beispielsituation dennoch darzustellen, wurden für den Fall  $H_0 : \sigma_a^2 = 0$  mittels 100.000 Simulationsläufen  $Z$ -Statistiken der Form

$$Z = \frac{\hat{\sigma}_a^2}{\text{SE}(\hat{\sigma}_a^2)} \quad (2.16)$$

gebildet und deren empirische Verteilung ermittelt. Für den Fall dass  $\hat{\sigma}_a^2 = 0$ , bei dem kein Standardfehler für diese Schätzung existiert, wurde  $Z = 0$  gesetzt. Die so gewonnene empirische Dichte und Verteilung sind in Abbildung 2.9 dargestellt.

**Abbildung 2.9: Simulationsergebnisse zu Z-Statistiken im Random-Logit-Modell unter  $\sigma_a^2 = 0$ , unbalanciert ( $N = 3.465$ ,  $p = 10$ ,  $\tilde{p} = 0,1146$ )**



Durch Betrachtung der empirischen (Zähl-)Dichte zeigt sich, dass die (positiv semi-definite) Verteilung der Z-Statistiken offensichtlich keiner (gestutzten) Standard-Normalverteilung entstammt. Für die Bestimmung der kritischen Grenzen  $c_\alpha$  für einen Niveau- $\alpha$ -Test auf den Streuungsparameter  $\sigma_a^2$  wird somit auf die Verwendung der Quantile der beobachteten empirischen Verteilung der Z-Statistiken in der hier gewählten Beispielsituation zurückgegriffen. Die folgenden Grenzen, die hiermit für die Anwendung am Beispieldatensatz empfohlen werden, wurden für ausgewählte einseitige Signifikanzniveaus wie folgt ermittelt:

$\alpha$ (einseitig)	kritische Grenze $c_\alpha$	$\alpha$ (einseitig)	kritische Grenze $c_\alpha$
0,200	0,5389	0,025	1,0779
0,100	0,7916	0,010	1,1980
0,050	0,9596	0,005	1,2711

Für den Fall  $H_1 : \sigma_a^2 \neq 0$  soll für verschiedene Werte von  $\sigma_a^2$  geprüft werden, wie oft die Nullhypothese zum einseitigen Niveau  $\alpha = 5\%$  mittels der oben beschriebenen kritischen Grenzen (d.h.  $c_{0,05} \approx 0,9596$ ) verworfen wird. Die Ergebnisse – basierend auf 10.000 Simulationen je Wert für  $\sigma_a^2$  – für die unbalancierte Situation des Anwendungsbeispiels sind in folgender Tabelle dargestellt.

$\sigma_a^2$	empirische Testgüte ( $\alpha = 0,05$ )	$\sigma_a^2$	empirische Testgüte ( $\alpha = 0,05$ )
0,00	0,0492	0,06	0,7773
0,01	0,1905	0,08	0,8634
0,02	0,3556	0,10	0,9154
0,03	0,5041	0,12	0,9392
0,04	0,6200	0,15	0,9674
0,05	0,7129	0,20	0,9844

#### 2.4.2.4 Eigenschaften von Zentrums-spezifischen Konfidenzintervallen und Hypothesentests

Abschließend werden nun wieder Aussagen über einzelne Kliniken hinsichtlich ihres individuellen Behandlungserfolgs getroffen und entschieden, ob sich dieser signifikant vom Gesamtmittel (hier der Grundwahrscheinlichkeit) unterscheidet. Zu diesem Zweck werden im Anwendungsbeispiel die Ereigniswahrscheinlichkeiten der Kliniken entsprechend des vordefinierten  $\sigma_a^2$  im linearen Prediktor erzeugt. Dabei werden wieder die beiden Fälle  $H_0$  ( $\sigma_a^2 = 0$ , und somit  $Z\gamma = 0$ , d.h. dass alle Zentren dieselbe Ereigniswahrscheinlichkeit  $p_i = \tilde{p} \forall i$  besitzen) und  $H_1$  ( $\sigma_a^2 > 0$ ) betrachtet.

Für den Signifikanztest zum Streuungsparameter  $\sigma_a^2$  wurde ein einseitiges Signifikanzniveau von 5% gewählt (vgl. Abschnitt 2.4.2.3). Nur im Falle der – unter  $H_0$  fälschlichen und unter  $H_1$  korrekten – Ablehnung von  $H_0$  wurden dann zweiseitige 95%-Konfidenzintervalle für  $Z\hat{\gamma} = \hat{\gamma}_i = \hat{a}_i$  betrachtet und mit dem Wert 0 verglichen.



Hypothesentests zu  $H_{0i}$  darzustellen. In der Darstellung wurden die Zentrumsnummern aus dem Beispieldatensatz (siehe Kapitel 4) benutzt und nach der Fallzahl  $n_i$  aufsteigend sortiert. Für den Fall der Alternativhypothese wurden exemplarisch die beiden Zentren mit der geringsten und höchsten Patientenzahl ausgewählt.

**Tabelle 2.12: Simulationsergebnisse zu Klinik-spezifischen Hypothesentests**  
( $\alpha = 0,05$ ), unbalanciert ( $p = 10$ ,  $N = 3.465$ ),  $\sigma_a^2 = 0$ ,  $\tilde{p} = p_i = 0,1146$

Klinik Nr.	Anteil an allen Versuchen			Anteil nach Entsch. „ $H_1$ “ *		
	(-1)	(0)	(+1)	(-1)	(0)	(+1)
7 (n=200)	0,019%	4,941%	0,063%	0,378%	98,368%	1,254%
6 (n=209)	0,014%	4,949%	0,060%	0,279%	98,527%	1,195%
4 (n=223)	0,031%	4,914%	0,078%	0,617%	97,830%	1,553%
10 (n=233)	0,027%	4,933%	0,063%	0,538%	98,208%	1,254%
9 (n=267)	0,054%	4,848%	0,121%	1,075%	96,516%	2,409%
8 (n=338)	0,061%	4,856%	0,106%	1,214%	96,675%	2,110%
1 (n=342)	0,052%	4,848%	0,123%	1,035%	96,516%	2,449%
3 (n=454)	0,097%	4,732%	0,194%	1,931%	94,207%	3,862%
5 (n=560)	0,143%	4,649%	0,231%	2,847%	92,554%	4,599%
2 (n=639)	0,134%	4,635%	0,254%	2,668%	92,276%	5,057%

\* Testgüte: 5,023%

**Nullhypothese:** Im Falle, dass  $\sigma_a^2 = 0$  (siehe Tabelle 2.12), sind in 5,02% der 100.000 Beispielstudien mittels der empirisch gefundenen kritischen Grenze (für  $\alpha = 0,05$  mit  $c_\alpha = 0,9596$ , vgl. Kapitel 2.4.2.3) Unterschiede zwischen den Kliniken aufgedeckt worden, obwohl keine Unterschiede vorliegen. Es zeigt sich, dass für die kleineren Zentren deutlich häufiger auf  $\gamma_i > 0$  als auf  $\gamma_i < 0$  geschlossen wurde. Bei den größeren Zentren ist dieser Unterschied schwächer ausgeprägt, aber dennoch markant.

**Alternativhypothese:** Für  $\sigma_a^2 = 0,05$  wurden in gut 71% aller Beispielauswertungen durch Betrachtung von  $\hat{\sigma}_a^2$ , dividiert durch den Standardfehler, Unterschiede zwischen den Zentren festgestellt. Die Einrichtung Nr. 2 mit  $n_2 = 639$  würde, falls in Wahrheit ein Effekt von  $\gamma_i \approx 0,3$  (d.h.  $p_i \approx 0,1487$ ) vorliegt, in ca. 40,7% jener Versuche, bei denen global auf  $\sigma_a^2 > 0$  geschlossen wurde, korrekterweise als in der Ereigniswahrscheinlichkeit über der Grundwahrscheinlich-

**Tabelle 2.13: Simulationsergebnisse zu Klinik-spezifischen Hypothesentests**  
 $(\alpha = 0,05)$ , unbalanciert ( $p = 10, N = 3.465$ ),  $\sigma_a^2 = 0,05, \tilde{p} = 0,1146$

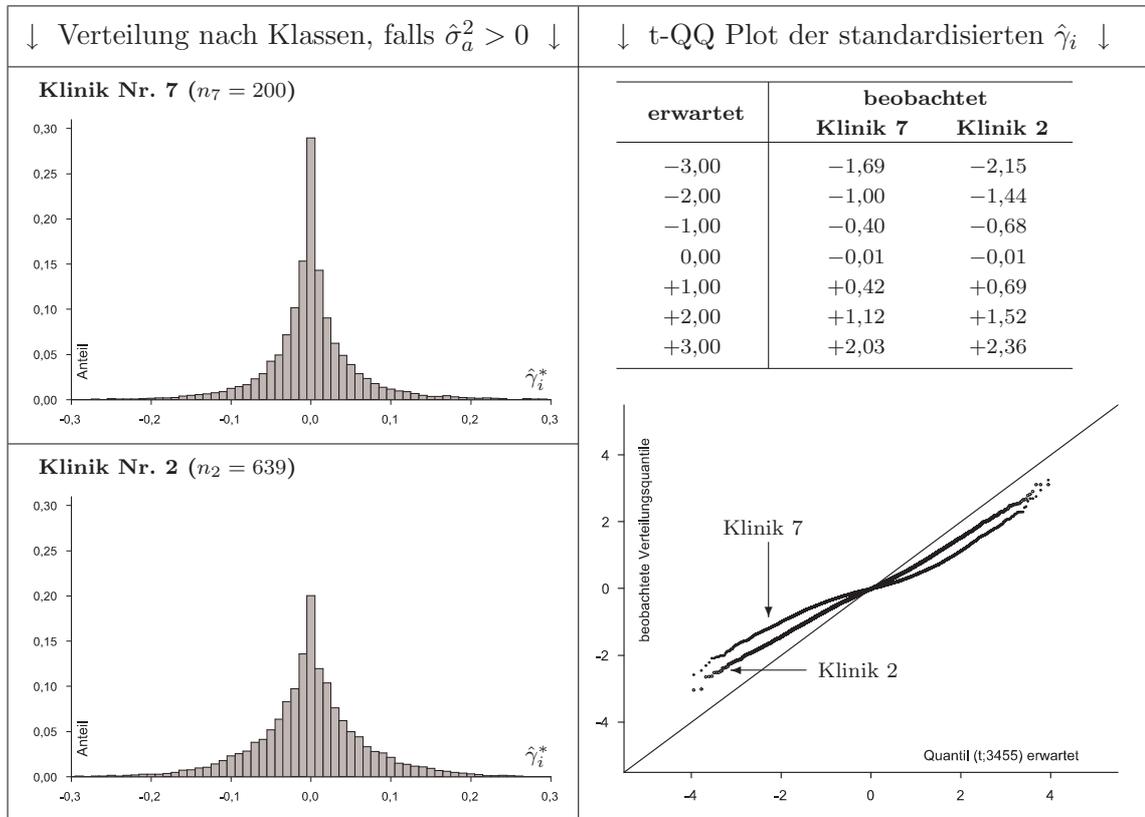
Klinik Nr.	$\gamma_i$ $\pm 0,05$	$p_i \approx$	$h(p_i)$	Anteil an allen Versuchen			Anteil nach Entsch. „ $H_1$ “ *		
				(-1)	(0)	(+1)	(-1)	(0)	(+1)
7 (n=200)	-0,6	0,0663	0,0051	20,497%	61,077%	0,000%	25,127%	74,873%	0,000%
	-0,5	0,0728	0,0151	12,759%	63,059%	0,000%	16,828%	83,172%	0,000%
	-0,4	0,0798	0,0367	6,788%	68,178%	0,000%	9,055%	90,945%	0,000%
	-0,3	0,0875	0,0730	3,174%	68,550%	0,000%	4,425%	95,575%	0,000%
	-0,2	0,0958	0,1194	1,580%	69,725%	0,025%	2,215%	97,750%	0,035%
	-0,1	0,1048	0,1604	0,696%	68,614%	0,127%	1,003%	98,815%	0,182%
	0,0	0,1146	0,1769	0,170%	69,116%	0,339%	0,244%	99,269%	0,487%
	0,1	0,1251	0,1604	0,043%	68,195%	1,221%	0,062%	98,180%	1,757%
	0,2	0,1365	0,1194	0,017%	67,224%	3,804%	0,024%	94,622%	5,354%
	0,3	0,1487	0,0730	0,000%	64,484%	9,168%	0,000%	87,552%	12,448%
	0,4	0,1618	0,0367	0,000%	60,362%	18,496%	0,000%	76,545%	23,455%
0,5	0,1758	0,0151	0,000%	49,705%	33,464%	0,000%	59,764%	40,236%	
0,6	0,1908	0,0051	0,000%	33,134%	55,689%	0,000%	37,303%	62,697%	
2 (n=639)	-0,6	0,0663	0,0051	80,354%	14,735%	0,000%	84,504%	15,496%	0,000%
	-0,5	0,0728	0,0151	63,030%	29,653%	0,000%	68,006%	31,994%	0,000%
	-0,4	0,0798	0,0367	41,131%	43,930%	0,000%	48,355%	51,645%	0,000%
	-0,3	0,0875	0,0730	24,772%	55,486%	0,000%	30,866%	69,134%	0,000%
	-0,2	0,0958	0,1194	10,666%	61,661%	0,017%	14,744%	85,233%	0,023%
	-0,1	0,1048	0,1604	3,231%	61,872%	0,087%	4,956%	94,911%	0,133%
	0,0	0,1146	0,1769	0,787%	60,757%	0,860%	1,260%	97,361%	1,378%
	0,1	0,1251	0,1604	0,113%	60,676%	3,892%	0,174%	93,808%	6,018%
	0,2	0,1365	0,1194	0,008%	58,374%	12,753%	0,012%	82,060%	17,928%
	0,3	0,1487	0,0730	0,000%	48,251%	33,098%	0,000%	59,313%	40,687%
	0,4	0,1618	0,0367	0,000%	29,951%	61,150%	0,000%	32,877%	67,123%
0,5	0,1758	0,0151	0,000%	15,431%	80,494%	0,000%	16,086%	83,914%	
0,6	0,1908	0,0051	0,000%	2,907%	96,318%	0,000%	2,930%	97,070%	

\* Testgüte: 71,47%

keit liegend klassifiziert werden. Im Gegensatz hierzu wäre die entsprechende Wahrscheinlichkeit für das kleinste Zentrum in der Untersuchung (Nr. 7,  $n_7 = 200$ ) lediglich 12,4%.

Der Effekt der bezüglich des linearen Prediktors unsymmetrischen Testentscheidungen ist in der limitierten Fallzahl und der damit verbundenen diskreten Eigenschaft der beobachtbaren Klinikeinflüsse, verbunden mit der relativen Randlage von

**Abbildung 2.10: Verteilung von EBLUP-Schätzern im Random-Logit-Modell,  $\sigma_a^2 = 0$ , unbalanciert ( $N = 3.465$ ,  $p = 10$ ,  $\tilde{p} = 0,1146$ )**

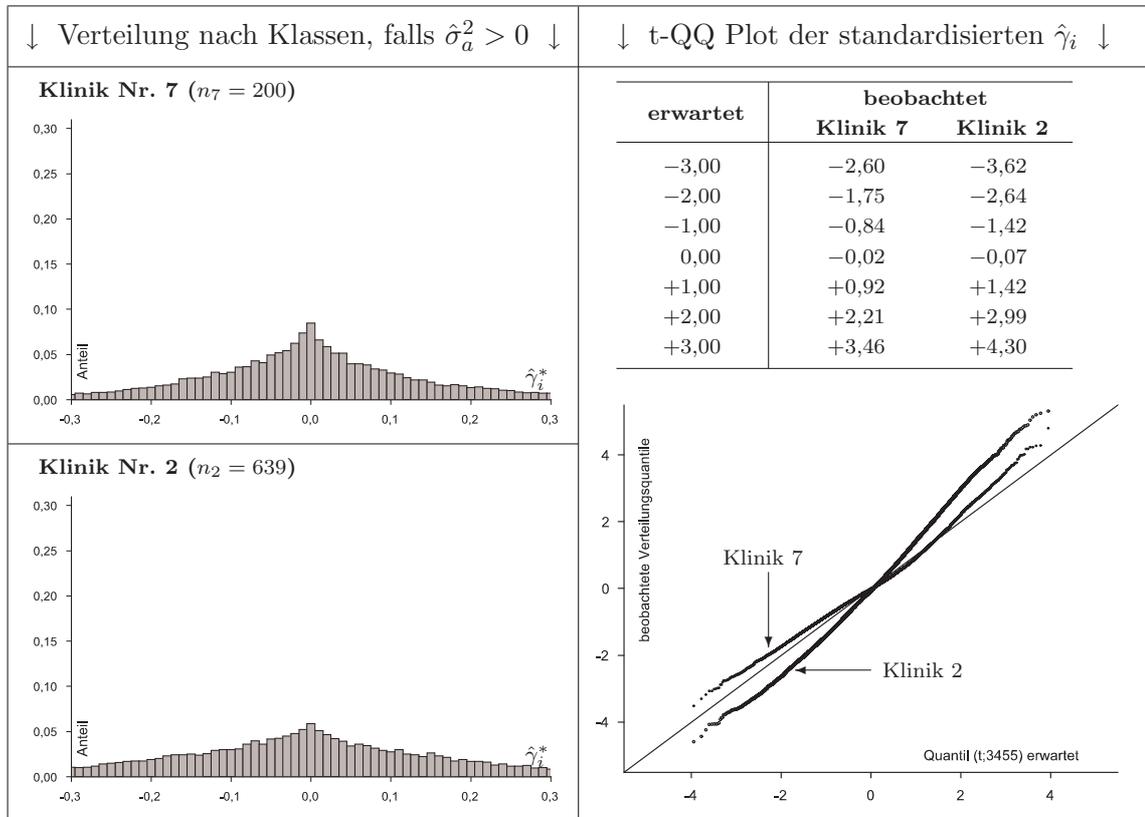


\* $\hat{\gamma}_i$  mit Klassenbreite = 0,01

$\tilde{p}$  im Wahrscheinlichkeitsraum, begründet. Da  $\tilde{p} < 0,5$  ist, weist jede Binomialverteilung  $B_i(p_i)$  für  $p_i < 0,5$  eine Linksschiefe und eine kleinere Anzahl von möglichen Ausprägungen unterhalb von  $E(y_i) = n_i p_i$  auf. Je größer die Fallzahl  $n_i$  eines Zentrums ist, desto geringer wirken sich die Schiefe und die Diskretheit aus. Die Verteilung der  $\hat{\gamma}_i$  ist somit – besonders für kleinere  $n_i$  – trotz eines Mittelwertes von 0 leicht rechtsschief, und dies für beliebige Werte von  $\sigma_a^2$ . Der Effekt der Varianzaufblähung hin zu größeren Ereignishäufigkeiten ist in den Tabellen 2.10 und 2.11 illustriert. Lässt man  $p_i$  gegen 0,5 oder  $n_i$  gegen  $\infty$  streben, ergeben sich Symmetrien hinsichtlich der Binomialverteilungen und damit der Testentscheidungen für die  $H_{0i}$ .

Betrachtet man die Originalskala, so sind die Konfidenzintervalle unsymmetrisch und wegen  $\tilde{p} < 0,5$  für kleine  $\hat{p}_i$  enger als für große  $\hat{p}_i$ . Im Falle von  $\tilde{p} > 0,5$  verhält sich dies in umgekehrter Weise. Bei gleichem absoluten Unterschied in der tatsächli-

**Abbildung 2.11: Verteilung von EBLUP-Schätzern im Random-Logit-Modell,  $\sigma_a^2 = 0,05$ , unbalanciert ( $N = 3.465$ ,  $p = 10$ ,  $\tilde{p} = 0,1146$ )**



\* $\hat{\gamma}_i$  mit Klassenbreite = 0,01

chen Ereigniswahrscheinlichkeit eines Zentrums vom Gesamtmittel sind die Ablehnwahrscheinlichkeiten für  $H_{1i} : \gamma_i < 0$ , gegeben  $p_i - \tilde{p} = c_i$  (falls  $c_i < 0$ ), trotz der beobachteten Asymmetrie in der Verteilung der EBLUP-Schätzer häufig höher als die entsprechenden Ablehnwahrscheinlichkeiten für  $H_{1i} : \gamma_i > 0$ , für  $c_i > 0$  (mit  $|c_i| = |c_{i'}|$ ).

Die empirisch gefundene Verteilung der  $\hat{\gamma}_i$ , deren leichte Asymmetrie zu den asymmetrischen Entscheidungen zu den Einrichtungs-spezifischen Hypothesentests  $H_{0i}$  führt, ist für das kleinste und das größte teilnehmende Zentrum (Nr. 7 und 2) in den Abbildungen 2.10 und 2.11 exemplarisch dargestellt, jeweils für  $\sigma_a^2 = 0$  und für  $\sigma_a^2 = 0,05$ . Alle Situationen, in denen  $\hat{\sigma}_a^2 = 0$ , und somit für alle  $\hat{\gamma}_i = 0$  gilt, wurden aus der Darstellung eliminiert, da diese keinen Informationsbeitrag zur Symmetrie der Verteilung liefern. Die Betrachtung der Verteilungen der standardisierten EBLUP-Schätzer (Werte der t-Statistiken, t-QQ-Plots) zeigt die Asymmetrie und

die nach Fallzahl unterschiedliche Form der Verteilungen. Für jede der beiden Kliniken wurden jeweils 25.000 Beobachtungen generiert.

## 2.5 Alternative Analyseverfahren klassischer Modelle

Üblicherweise wird in der schließenden Statistik im Falle verletzter Verteilungsannahmen auf nicht-parametrische bzw. verteilungsfreie oder auch exakte Verfahren ausgewichen. Die Testergebnisse beispielsweise bei Gruppenvergleichen (als feste Effekte modelliert) sind gegenüber den Verteilungsannahmen robust, besitzen jedoch häufig eine schwächere Güte als die entsprechenden parametrischen Tests, insbesondere dann, wenn die Verteilungsannahmen tatsächlich nicht nachhaltig verletzt sind.

Beim Vorliegen zufälliger Effekte und dem damit verbundenen zu wählenden Ansatz gemischter Modelle sowie bei Klumpen- bzw. hierarchischen Datenstrukturen sind bisher kaum nicht-parametrische Verfahren bekannt. Insbesondere werden von den gängigen Software-Paketen bisher keine derartigen Lösungen angeboten, so dass im Einzelfall mit selbst konstruierten Algorithmen gearbeitet werden müsste.

Zwar können einerseits bei Einrichtungsvergleichen die im linearen Modell geforderten Annahmen zur Normalverteilung der Streuungsursachen (Verteilung der Zentrumseffekte und der Patienteneffekte) durchaus angezweifelt werden, andererseits sind die zugehörigen Datenumfänge meist recht groß ( $N > 1.000$ ), so dass zumindest die asymptotischen Verteilungsannahmen für die verwendeten Teststatistiken in der Regel als erfüllt angesehen werden können.

## 2.6 CART Verfahren

### 2.6.1 Motivation

Das Verfahren der Klassifikations- und Regressionsbäume „CART“ geht zurück auf Breiman et al. (1984) [4].

Bei diesem Ansatz werden Zusammenhänge zwischen den betrachteten (erhobenen) Einflussgrößen und der Zielgröße mittels Entscheidungsregeln dargestellt. Hierzu werden die Subjekte (z.B. Patienten) als Teil der Gesamtpopulation entsprechend ihrer (klassifizierten) Merkmale/Einflussgrößen mittels eines rekursiven Partitionierungs-Algorithmus in Untergruppen eingeteilt, so dass die Zusammenhänge im Sinne von Entscheidungsregeln bzw. logischen Verknüpfungen der Form

„Falls für die Einflussgrößen Eigenschaft A gilt,  
dann gilt für die Zielgröße Eigenschaft B“

dargestellt werden können.

Dazu Breiman et al.[4] (S. 23):

*„The fundamental idea is to select each split of a subset so that the data in each of the descendant subsets are “purer“ than the other data in the parent subset“.*

Die Bezeichnung als „Klassifikations-“ oder „Regressionsbaum“ hängt vom Skalenniveau der Zielgröße ab. Liegt die Zielgröße in ordinal- oder nominal-skaliertem Form vor, so spricht man von **Klassifikationsbäumen**; bei stetigen Zielgrößen werden **Regressionsbäume** betrachtet.

Aufgrund der Gruppeneinteilung müssen die Einflussgrößen bei Klassifikations- und Regressionsbäumen in klassifizierter Form vorliegen (wie etwa Geschlecht oder Raucherstatus). Hier liegt die Gruppeneinteilung bereits aufgrund der Ausprägungen vor, wogegen bei stetigen Größen noch geeignete Klassifizierungen gefunden werden müssen. Das Alter als Einflussgröße kann zu diesem Zweck in geeignete Klassen, beispielsweise in *jung* ( $\leq 60$  Jahre) bzw. *alt* ( $> 60$  Jahre), zerlegt werden.

Bei jeder Gruppeneinteilung muss jedes Subjekt aus der Population bzw. jeweiligen Untergruppe eindeutig zuzuordnen sein, d.h. die zugrunde liegende Menge wird **vollständig** in **disjunkte** Teilmengen zerlegt („disjunkte Partitionierung“).

Beispiel für eine einfache Entscheidungsregel: Sollte das Geschlecht eines Patienten als relevanter Einflussfaktor für die Wahrscheinlichkeit des 30-Tage-Überlebens (ja/nein) eines Herzinfarktpatienten gelten, so könnte eine Entscheidungsregel lauten:

- falls Geschlecht = *männlich*, dann gilt für Zielgröße  $\hat{p} = 40\%$ ;
- falls Geschlecht = *weiblich*, dann gilt für Zielgröße  $\hat{p} = 60\%$ ,

wobei  $p$  die Wahrscheinlichkeit für das Überleben nach 30 Tagen bezeichnet. Für alle Versuchseinheiten (z.B. Patienten) innerhalb eines Knotens wird dann diejenige Ausprägung von  $Y$  als „Vorhersagewert“ bezeichnet, die innerhalb dieses Knotens die höchste Häufigkeit besitzt. Die Entsprechung des Vorhersagewerts ist bei Regressionsbäumen eine Parameterschätzung.

Ein Klassifikations- und Regressionsbaum hat die in Abbildung 2.12 dargestellte Struktur.

In diesem Beispiel wird die Gesamtpopulation im ersten Schritt nach einer zweistufigen Variablen  $X_1$  unterteilt (Bedingung 1, z.B. „Geschlecht = *männlich*“).

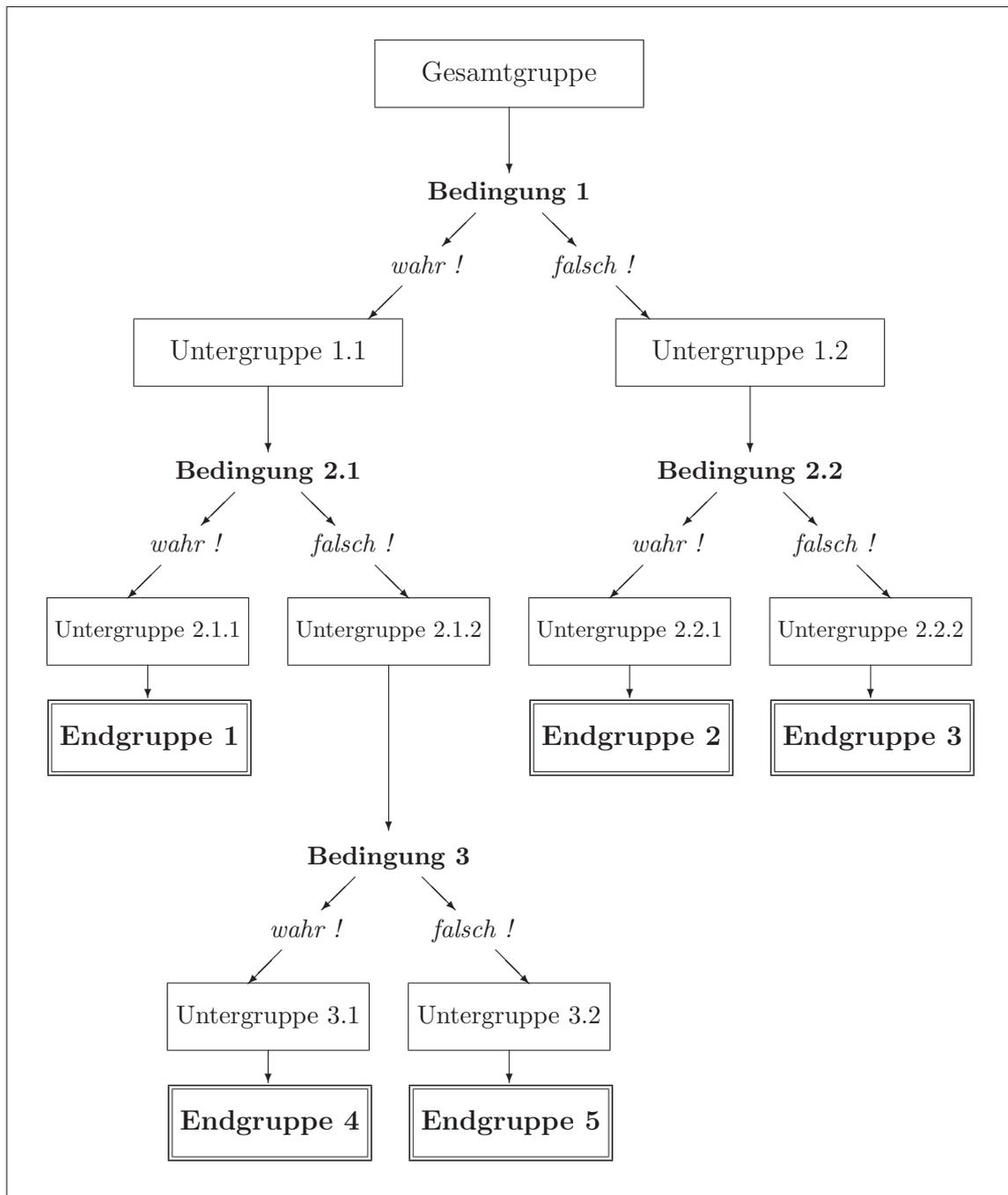
Somit ergeben sich nach der ersten Stufe zwei Untergruppen:

- Untergruppe 1.1 bezeichnet die Männer;
- Untergruppe 1.2 bezeichnet die Frauen.

Der Klassifikationsalgorithmus stellt für beide Untergruppen 1.1 und 1.2 weitere relevante Klassifikationsvariablen fest, so dass beide Gruppen weiter zerteilt werden. Im zweiten Schritt wird somit – für beide Untergruppen getrennt – wieder nach einer zweistufigen Variablen  $X_2$  unterteilt (Bedingung 2, z.B. „Rauchstatus = *Raucher*“). D.h.,

- Untergruppe 2.1.1: männliche Raucher;

Abbildung 2.12: Schematische Darstellung eines Klassifikations- bzw. Regressionsbaums



- Untergruppe 2.1.2: männliche Nichtraucher;
- Untergruppe 2.2.1: weibliche Raucher;

- Untergruppe 2.2.2: weibliche Nichtraucher.

Nun wird festgestellt, dass für die Untergruppen 2.1.1, 2.2.1 und 2.2.2 keine weiteren im Modell betrachteten Variablen mehr für die Zielgröße relevante Information liefern. Die genannten drei Untergruppen werden infolgedessen zu den ersten drei **Endgruppen**:

- Endgruppe 1: männliche Raucher;
- Endgruppe 2: weibliche Raucher;
- Endgruppe 3: weibliche Nichtraucher.

Für die Gruppe 2.1.2 (männliche Nichtraucher) spielt jedoch noch eine weitere Variable hinsichtlich der Überlebenswahrscheinlichkeit eine Rolle, nämlich das Alter. Angenommen, das Alter sei, wie oben beschrieben, in die Klassen in *jung* ( $\leq 60$  Jahre) bzw. *alt* ( $> 60$  Jahre), zerlegt worden. Damit lässt sich auch für die nicht-stetige Einflussgröße  $X_3$  (*Alter*) eine binäre Bedingung (Bedingung 3, z.B. „Alter = *alt*“) aufstellen. Entsprechend ergeben sich in der dritten Stufe noch zwei Untergruppen:

- Untergruppe 3.1: männliche Nichtraucher, alt;
- Untergruppe 3.2: männliche Nichtraucher, jung.

Nun seien auch für diese beiden Untergruppen keine weiteren für die Zielgröße relevanten Variablen mehr festzustellen. Die beiden Untergruppen 3.1 und 3.2 werden infolgedessen zu **Endgruppen**,

- Endgruppe 4: männliche Nichtraucher, alt;
- Endgruppe 5: männliche Nichtraucher, jung,

womit der Entscheidungsbaum mit insgesamt fünf disjunkten Endgruppen abgeschlossen ist. Für jede der gefundenen Endgruppen können nun Schätzungen für die Zielgröße angegeben werden. Die Unter- bzw. Endgruppen werden im Zusammenhang mit CART-Analysen auch als **Knoten** bezeichnet.

Je nach Datenlage, nach der Zahl der betrachteten Einflussfaktoren und der Wahl der Entscheidungsregeln kann ein solcher Entscheidungsbaum natürlich wesentlich komplexere Gestalt annehmen.

In dem beschriebenen Beispiel wurden die beiden Untergruppen 1.1 und 1.2 in Stufe zwei mit ein und derselben Variablen weitergeführt. Dies muss jedoch nicht notwendigerweise so sein. Es ist ebenfalls denkbar, dass für die Männer (Gruppe 1.1) z.B. der Rauchstatus als Bedingung 2.1 und für den Frauen der Alkoholkonsum als Bedingung 2.2 formuliert wird.

Wie bereits ausgeführt, lassen sich nicht nur nominal skalierte oder binäre – wie im obigen Beispiel – sondern auch ordinal skalierte Variablen in Klassifikations- und Regressionsbäumen modellieren. Liegt ein Einflussfaktor beispielsweise in  $k$  Ausprägungsstufen (z.B. die Schulbildung in  $k = 4$  Stufen, als *kein Abschluss*, *Hauptschule*, *Realschule* und *Gymnasium*) vor, so ergeben sich für eine solche Partitionierung entsprechend  $k$  Untergruppen.

Allerdings sind bei CART-Analysen in der Praxis nur binäre Splits der Form  $X_i \leq c_i$  erlaubt, was die Definition von geeigneten Dichotomisierungen nötig macht. Bei  $k$  Ausprägungen einer ordinal skalierten Variablen sind aufgrund ihrer inhärenten Ordnung  $k - 1$  Dichotomisierungen möglich. Bei nominal skalierten (rein kategoriellen) Einflussgrößen (ohne Ordnung) mit  $k > 2$  Ausprägungen (wie zum Beispiel Stadt bzw. Stadtteil) sind  $2^{k-1} - 1$  (entsprechend der um eins reduzierten linken oder rechten Hälfte der  $k + 1$ 'ten Zeile im Pascal'schen Dreieck) möglich. Bei  $k = 4$  Ausprägungen wären dies,  $2^3 - 1 = 7$ , bei  $k = 5$  bereits  $2^4 - 1 = 15$  Aufteilungen (vgl. Abschnitt 2.6.3.1). Bei kontinuierlichen Kovariaten sind theoretisch beliebig viele Aufteilungen möglich.

Es zeigt sich, dass für eine möglichst einfache Baumstruktur binäre oder ordinale Skalenniveaus der Einflussfaktoren am besten geeignet sind.

Durch die Partitionierung der Gesamtpopulation in disjunkte (End-)Gruppen sollte eine möglichst scharfe Trennung hinsichtlich der Zielgröße erreicht werden, d.h. die Entscheidungsregel

$$\text{Falls } X_i = x_i \text{ (} i = 1, \dots, p \text{), dann gilt } Y = y$$

sollte möglichst allgemeingültig sein. Hiermit ist in der Praxis jedoch meist nicht zu rechnen, da ein Modell niemals *alle* möglichen Einflüsse einbeziehen kann (und auch nicht sollte), wodurch immer ein individueller Zufallsaspekt bleibt (entsprechend dem Restfehler  $e$  bei linearen Modellen). Vielmehr werden die Zusammenhänge in Form von Verteilungen (im Fall von Regressionsbäumen) bzw. Wahrscheinlichkeiten (bei Klassifikationsbäumen) angegeben. Das bedeutet, dass sich auch bei starken Assoziationsstrukturen Überlappungen zwischen den Gruppen ergeben. Je stärker der Zusammenhang ist, desto geringer wird allerdings diese Überlappung ausfallen.

### 2.6.2 Modellbildung

In diesem Abschnitt gilt die folgende Notation:

$X_i$	$i$ -te Einflussgröße ( $i = 1, \dots, k$ )
$k$	Anzahl der modellierten Einflussgrößen $X_i$
$x_i$	Ausprägung der $i$ -ten Einflussgröße $X_i$
$Y$	Zielgröße
$y$	Ausprägung der Zielgröße $Y$
$p$	Anzahl möglicher Ausprägungen einer kategoriellen Zielgröße $X_i$ (die Ausprägungen seien hierbei stets von $1, \dots, p$ nummeriert)
$c_i$	Aufteilungspunkt („split point“) einer nicht-binären Einflussgröße $X_i$
$g$	„Knoten“ (Ausgangs-, Unter- oder Endgruppe)
$n_g$	Anzahl Beobachtungen im Knoten $g$
$g_L$	linker Tochterknoten
$n_L$	Anzahl Beobachtungen im linken Tochterknoten $g_L$
$g_R$	rechter Tochterknoten
$n_R$	Anzahl Beobachtungen im rechten Tochterknoten $g_R$
$\hat{p}(i g)$	Anteilsschätzer für kategorielle Zielgröße („Vorhersagewert“ für Ausprägung $i \in \{1, \dots, m\}$ , gegeben einen Knoten $g$ )
$\pi(g_j)$	Fallzahlanteil im Tochterknoten $g_j$ am Knoten $g$
$\phi(c_i, g)$	Split Funktion für stetige Zielgröße
$\Omega_g$	Gesamtmenge aller erlaubten Splits in $g$
$SS(g)$	Fehlerquadratsumme (Sum of Squares) in Knoten $g$
$i(g)$	(Un)Reinheits- („(Im)Purity“) Index
$\Delta(c_i, g)$	Split Funktion für diskrete Zielgröße

Im Bereich der Versorgungsforschung werden bei Registerdaten in der Regel zahlreiche Einflussfaktoren erhoben. Diese umfassen zumeist soziodemographische, anamnestiche und geographische Faktoren. Die Auswahl der in die Analyse einzubeziehenden Kovariaten sollte – genau wie bei den linearen Modellen – in Absprache mit den zuständigen Substanzwissenschaftlern geschehen. Da die Assoziationsstrukturen vor der Analyse nicht bekannt sind oder nur vermutet werden können, sollten zunächst aber eher eine zu große Zahl von Faktoren einbezogen werden als eine zu geringe. Die Durchführung einer Regressions- oder Klassifikationsbaum-Schätzung

wird somit auf Basis aller Beobachtungen der Zielgröße und aller (auch nur möglicherweise) mit ihr assoziierten Einflussgrößen durchgeführt.

Als Einschränkung hierzu sei gesagt, dass die Subjekt Daten zur Durchführung hinsichtlich aller ins Modell einbezogenen Faktoren vollständig sein müssen, damit das jeweilige Subjekt in die Analyse einbezogen werden kann. Falls die Datenlage lückenhaft ist und sollten Vorüberlegungen darauf schließen lassen, dass bestimmte Variablen hinsichtlich ihres Einflusses auf das Zielkriterium von geringerem Interesse sind, kann auf dieser Basis eine Vorab-Variablenselektion durchgeführt werden, um die Datenbasis möglichst umfänglich zu halten.

Alternativ kann jedoch eine Kategorie „fehlend“ als zulässige Ausprägung definiert werden, wodurch theoretisch alle Subjekte einbezogen werden können. Allerdings sollte zuvor eine Untersuchung auf zufällig oder systematisch fehlende Werte geschehen, um die Datenlage besser verstehen zu können. Auf weitere Aspekte der Modellbildung wird im Rahmen dieser Arbeit nicht eingegangen.

### 2.6.3 Modellschätzung

Die Schätzung der Parameter im Klassifikations- bzw. Regressionsbaums erfolgt, wie in der Einführung ausgeführt, über eine disjunkte Zerlegung der Gesamtgruppe nach relevanten Einflussfaktoren. Der Aufteilungsmechanismus erfolgt dabei rekursiv (siehe Hand [18]) unter Einbeziehung zunächst aller verfügbaren Einflussgrößen.

Ausgehend von der unpartitionierten Menge (Gesamtpopulation) werden zunächst für jeden Einflussfaktor  $X_i$  ( $i = 1, \dots, p$ ) **einzeln** alle (falls mehr als eine möglich ist) möglichen Aufteilungen betrachtet und miteinander verglichen. Die hierbei jeweils „beste“ Aufteilung (zur Definition siehe folgende Abschnitte) wird festgehalten.

Mittels dieser  $p$  jeweils besten Aufteilungen wird nun die insgesamt „beste“ Aufteilung ermittelt. Somit ergibt sich eine erste Bedingung, von der aus sukzessive – rekursiv – weitere Aufteilungskriterien ermittelt werden, so lange bis bestimmte „Stopp-Bedingungen“ erfüllt sind.

Die Wahl des Aufteilungsmechanismus hängt von den folgenden Kriterien ab:

- das **Split-Kriterium** (Aufteilungskriterium), welches angibt, wie die Aufteilungen hinsichtlich ihrer Optimalität beurteilt werden;

- die **Stopp-Regel** (Endkriterium) gibt die Bedingung an, ab wann in einer bestimmten Untergruppe keine weitere Aufteilung mehr erfolgen soll, d.h. wann eine Untergruppe zu einer Endgruppe wird (vgl. Abbildung 2.12). Die geeignete Größe (der Verzweigungsgrad) des Baums wird damit bestimmt;
- die Bestimmung von **Schätz- bzw. Vorhersagewerten** für die Zielgröße für die jeweiligen Endgruppen. Hierbei ist auf die Trennschärfe bzw. die Schätzgenauigkeit zu achten.

Die Bestimmung der drei genannten Merkmale wird in den folgenden Unterabschnitten beschrieben.

### 2.6.3.1 Bestimmung von Split-Regeln

Für die Beurteilung von Aufteilungen (Splits) müssen zunächst Grundregeln für die Zulässigkeit von Splits formalisiert werden. Everitt et al. ([11], S.103 ff.) definieren erlaubte Splits wie folgt:

1. jede Aufteilung basiert auf einer **einzelnen** Kovariaten  $X_i$ ;
2. für mindestens ordinal skalierte (geordnet kategorielle oder kontinuierliche) Variablen  $X_i$  muss eine Bedingung (ein Split)  $c_i$  der Form

$$g_L = \{X|X_i \leq c_i\} \quad \text{und} \quad g_R = \{X|X_i > c_i\} \quad (2.17)$$

für ein  $c_i$  aus dem Wertebereich von  $X_i$  aufgestellt werden. Die  $c_i$  heißen **split points** oder **Aufteilungspunkte**, durch die je Ausgangs- bzw. Untergruppe („Knoten“)  $g$  ein linker Tochterknoten  $g_L$  und ein rechter Tochterknoten  $g_R$  definiert wird;

3. für rein kategorielle Variablen sind alle möglichen disjunkten Aufteilungen der Gesamt- bzw. Untergruppe erlaubt.

Durch diese Kriterien sind gewisse Einschränkungen formuliert. Forderung 1 gibt vor, dass eine Aufteilung immer nur bezüglich einer Einflussgröße geschehen kann.

Das zweite Kriterium beschreibt eine disjunkte Aufteilung in zwei halboffene Intervalle. Wie bereits im vorigen Abschnitt angedeutet, gibt es bei ordinal skalierten Variablen  $k - 1$  und bei kontinuierlichen Variablen theoretisch überabzählbar viele Möglichkeiten für die Wahl von  $c_i$ . Wird jedoch vorgegeben, dass die  $c_i$  genau in der Mitte eines Intervalls zwischen zwei Werten in der geordneten Reihe aller Werte liegen sollen und betrachtet man stets endliche Stichprobengrößen, fällt die Zahl der relevanten Aufteilungsmöglichkeiten auf höchstens  $n - 1$  (nämlich dann, wenn die  $n$  Subjekte in  $X_i$  genau  $n$  unterschiedliche Werte tragen) zusammen.

Bei ungeordneten kategoriellen Variablen liegen, wie ebenfalls bereits ausgeführt, theoretisch  $2^{k-1} - 1$  mögliche Dichotomisierungen vor. Unter Einbeziehung von W.D. Fisher (1958) [13] reduzierte Breiman diese Zahl jedoch auf  $k$  mögliche Splits (siehe Breiman et al [4], Seite 101).

Wernecke et al. [60] definieren die Zielsetzung der Klassifikationsbäume wie folgt.

*„Given a set of risk factors  $X_1, \dots, X_k$  which influence a response variable  $Y$ , construct subgroups of all data which are internally as homogeneous and externally as heterogeneous as possible, measured on a characteristic function  $F(Y|X)$ .“*

Die Endgruppen sollen also so gewählt sein, dass sich die Subjekte in ihr bezüglich der Einflussgrößen möglichst wenig unterscheiden, dafür aber im Vergleich zu den anderen Endgruppen bezüglich der Zielgröße möglichst große (großer Abstand) und möglichst scharfe (möglichst keine Reststreuung) Unterschiede aufweisen.

Die verschiedenen in der Literatur diskutierten Split-Regeln unterscheiden sich im Wesentlichen durch die Fokussierung, einerseits auf die interne Homogenität oder andererseits die externe Heterogenität. Wie noch zu zeigen ist, stehen beide Ansätze durchaus in einem Spannungsverhältnis zueinander.

Für die Beurteilung aller möglichen und nach obiger Definition erlaubten Aufteilungen muss die geforderte Homogenität bzw. Inhomogenität zunächst definiert werden. Hierzu wird ein Maß für die (In)Homogenität eines bestimmten Knotens (für eine Bedingung) in Bezug auf die Zielgröße bestimmt.

Dieses Maß quantifiziert, wie stark durch die jeweilige Partitionierung die Inhomogenität der Gesamtgruppe im Vergleich zu der Summe der einzelnen Knoten reduziert wird.

**a) Stetige Zielgrößen**

Bei stetigen Zielgrößen wird die Inhomogenität innerhalb eines Knotens über die Devianz der Einzel-Beobachtungen

$$D(y_i; \hat{\mu}_i) = (y_i - \hat{\mu}_i)^2$$

definiert. Der Erwartungswert des Knotens, innerhalb dessen die Beobachtung  $y_i$  liegt, wird als Maximum-Likelihood-Schätzung mittels des arithmetischen Mittels  $\bar{y}_i$  bestimmt. Die Inhomogenität über die Summe der einzelnen Devianzen, also entsprechend den linearen Modellen, ist als Fehlerquadratsumme (Sum of Squares [SS]) bestimmt.

Die LS (Least Squares) Split-Funktion ist

$$\phi(c_i, g) = SS(g) - (SS(g_L) + SS(g_R)) \quad , \quad \text{mit} \quad SS(g) = \sum_{i \in g} (y_i - \bar{y})^2 .$$

Hierbei bezeichnet  $s$  einen bestimmten Split im Knoten  $g$ ,  $g_L$  und  $g_R$  beschreiben hier den linken bzw. rechten Tochterknoten (vgl. (2.17) auf Seite 121). Gesucht wird nun ein bestimmter Split  $c_i^*$ , für den gilt:

$$\phi(c_i^*, g) = \max_{c_i \in \Omega_g} \phi(c_i, g) \quad ,$$

wobei  $\Omega_g$  alle erlaubten Splits in  $g$  bezeichnet.

**b) Kategorielle Zielgrößen**

Bei diskreten Zielgrößen wird ein ähnlicher Ansatz gewählt. Die Split-Funktion

$$\Delta i(c_i, g) = i(g) - (\pi(g_L) i(g_L) + \pi(g_R) i(g_R))$$

beschreibt den Grad der Inhomogenitäts-Reduktion, gegeben durch den Split  $c_i$ , über einen „Unreinheits“- (Impurity) Index  $i(\cdot)$ . Die  $\pi(g_j)$  geben die Anteile der Subjekte an dem Gesamtknoten an, also  $\pi(g_j) = n_j/n_g$ ,  $j \in \{L; R\}$ .

Für  $i(\cdot)$  werden in der Literatur verschiedene Vorschläge diskutiert (vgl. Hand [18]).

1. Die **Fehlklassifikationsrate** gibt an, wie groß der Anteil der Subjekte in einem Knoten ist, die nicht die Kategorie mit der häufigsten Ausprägung zeigen:

$$i_1(g) = 1 - \max_{j=1,\dots,m} (\hat{p}(j|g)) ,$$

wobei  $\hat{p}$  den Anteilsschätzer in der  $j$ -ten Kategorie und  $m$  die Anzahl möglicher Ausprägungen von  $Y$  bezeichnen.

Im Falle von  $m = 2$  (binäre Zielgröße) gilt dann

$$\begin{aligned} i_1(g) &= 1 - \max\{\hat{p}(1|g) ; 1 - \hat{p}(1|g)\} \\ &= 1 - \max\{\hat{p}(2|g) ; 1 - \hat{p}(2|g)\} . \end{aligned}$$

2. Ein weiteres Maß, welches bereits von Breiman [4] besprochen wurde, ist der **Gini-Index**, der auch in vielen anderen Bereichen (wie beispielsweise zur Bestimmung der Ungleichverteilung von Einkommen) eingesetzt wird:

$$\begin{aligned} i_2(g) &= \sum_{i \neq j}^m \hat{p}(i|g) \hat{p}(j|g) \\ &= 2 \sum_{i=1}^m \sum_{j=i+1}^m \hat{p}(i|g) \hat{p}(j|g) . \end{aligned}$$

Bei einer binären Zielgröße vereinfacht sich die Schreibweise zu

$$\begin{aligned} i_2(g) &= 2 \hat{p}(1|g) (1 - \hat{p}(1|g)) \\ &= 2 \hat{p}(2|g) (1 - \hat{p}(2|g)) . \end{aligned}$$

3. Die **Entropie** oder „Deviance-Statistik“, motiviert durch die multinomiale Verteilung des gesamten Klassifikationsbaums, wird die (Un-)Reinheit definiert als

$$i_3(g) = - \sum_{i=1}^m \hat{p}(i|g) \log_b \hat{p}(i|g) .$$

Einige Überlegungen zur Basis  $b$  des Logarithmus-Terms folgt im folgenden Absatz. Bei  $m = 2$  gilt für  $i_3(g)$  entsprechend (mit  $b = 4$ , siehe unten stehende Nebenrechnung 3):

$$\begin{aligned} i_3(g) &= -\hat{p}(1|g) \log_4(\hat{p}(1|g)) - (1 - \hat{p}(1|g)) \log_4(1 - \hat{p}(1|g)) \\ &= -\hat{p}(2|g) \log_4(\hat{p}(2|g)) - (1 - \hat{p}(2|g)) \log_4(1 - \hat{p}(2|g)) . \end{aligned}$$

Weitere Überlegungen hinsichtlich der Eigenschaften (Minima, Maxima und Monotonität) der drei eingeführten Unreinheitsmaße werden im Anhang B.2 dargestellt.

Zusammenfassend kann zu den Eigenschaften Folgendes erklärt werden:

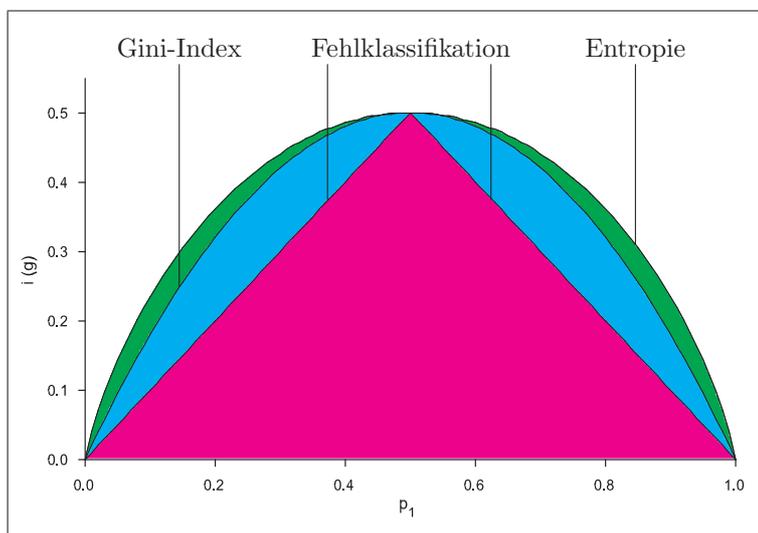
- Im Falle starker Konzentration eines kategoriellen Merkmals auf eine einzige Ausprägung wird die Inhomogenität (Impurity) minimal (gleich 0).
- Bei Gleichverteilung eines kategoriellen Merkmals, d.h.

$$\hat{p}(i|g) = \frac{1}{m}, \forall i = 1, \dots, m$$

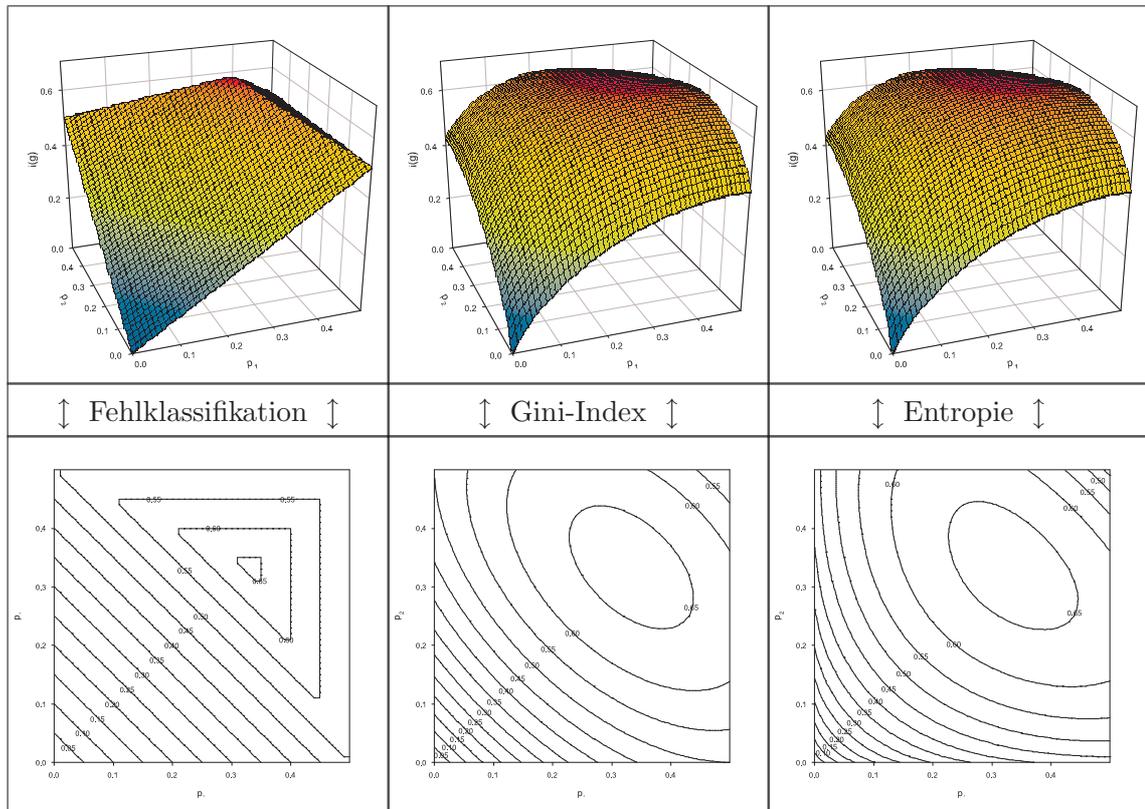
(„Laplace“-Wahrscheinlichkeiten als Schätzer), wird die Inhomogenität  $i(g)$  maximal, nämlich gerade gleich  $\frac{m-1}{m}$ . Mit wachsender Zahl der Ausprägungen strebt  $i(g)$  bei Gleichverteilung also gegen 1.

Einen Vergleich der drei eingeführten Impurity-Maße zeigt die Abbildung 2.13 am Beispiel von  $m = 2$  ( $p$  bezeichnet hier der Anteilsschätzer für einen beliebigen Tochterknoten).

**Abbildung 2.13: Impurity-Maße bei binärer Zielgröße**



Es zeigt sich, dass das Maß der Entropie im Vergleich zur Fehlklassifikation und zum Gini-Index – mit Ausnahme von  $p \in \{0; 0,5; 1\}$  – stets die größte Unreinheit ausweist. In den genannten drei Stützstellen sind die drei Maße identisch.

Abbildung 2.14: Impurity-Maße bei  $m = 3$ 

Die Abbildungen für zwei und drei Ausprägungen der Zielgröße verdeutlichen die Unterschiede zwischen den drei Maßen. Die Fehlklassifikationsrate ist mit einer Kostenfunktion vergleichbar, die jeder Zuweisung eines Subjekts zum „falschen“ Knoten dieselben Kosten zuweist. Beim Gini-Index und bei der Entropie werden den Fehlklassifikationen jedoch variable Kosten zugewiesen, weshalb sie auch „Variable Misclassification Costs“ genannt werden. Die Benutzung der beiden letzteren Reinheitsmaße führt dazu, dass besonders „reine“ Knoten bevorzugt werden, sofern dies möglich ist. Als häufig genutzter „Kompromiss“ hat sich der Gini-Index erwiesen.

Letztendlich ist es jedoch für die Schätzgenauigkeit des (finalen) Baums nicht erheblich, welches Maß benutzt wird, weshalb sich die Frage anschließt, warum überhaupt verschiedene Maße existieren. Man möchte zwar die Unreinheit innerhalb der Knoten minimieren, zum anderen möchte man aber vermeiden, dass Bäume zu viele Knoten aufweisen, um ihre Übersichtlichkeit und Interpretierbarkeit zu wahren. Außerdem sollte einer Variablen, die einen Split höherer Ordnung bewirkt, mehr

Interesse gewidmet werden als einer Einflussgröße, die spätere Splits auslösen. In einem Baum mit ungeeigneter Split-Regel wird nicht mehr gut ersichtlich sein, welche der Variablen wie stark für das Ergebnis entscheidend waren.

### Betrachtungen zur Modellgüte

Zur Beurteilung von Klassifikationsbäumen werden, entsprechend der Anpassungsgüte bei linearen Modellen (vgl. Seite 37), Bestimmtheitsmaße betrachtet.

#### 2.6.3.2 Bestimmung von Stopp-Regeln

Obwohl prinzipiell eine Baum-Partitionierung grundsätzlich soweit fortgeführt werden könnte, bis jede Endgruppe nur noch aus einzelnen Subjekten besteht, ist ein so gebildeter Klassifikations- oder Regressionsbaum nicht anzustreben. Zum einen würden so zu komplexe und kaum mehr interpretierbare Strukturen entstehen; zum anderen würden alle Subjekte gewissermaßen als Einflussfaktor betrachtet werden und somit kein – analog zum linearen Modell – zufälliger „Restfehler“ mehr zugelassen werden. Diese Problematik ist wiederum bei den linearen Modellen als „overfit“ bekannt. Daher ist es erforderlich, zu entscheiden, an welcher Stelle die Partitionierung zu beenden ist, und dem Algorithmus entsprechende Stopp-Kriterien vorzugeben.

Ein intuitiv verständlicher Ansatz ist es, als Stopp-Kriterium eine maximale Reduktion der Inhomogenität zu betrachten. Hierzu können die in Kapitel 2.6.3.1 beschriebenen Homogenitätsmaße herangezogen werden. Sollte durch einen nächsten Split die Inhomogenitäts-Reduktion nicht unter einem bestimmten Wert liegen, so müsste die weitere Partitionierung beendet werden.

Weiterhin ist denkbar, Maße für die Unterschiede zwischen den im jeweiligen Partitionierungsschritt zu bildenden Knoten zu betrachten, wie beispielsweise – bei stetigen Zielgrößen – durch die Feststellung von „signifikanten“ Lokationsunterschieden zwischen den Gruppen. Liegt der p-Wert einer t-Test- bzw. Wilcoxon-Test-Statistik unter einem vorgegebenem Wert für  $\alpha$ , so wäre der betrachtete Split durchzuführen, andernfalls nicht. Bei Pfeiffer et al. [44] wird zu diesem Vorschlag ein lokales Signifikanzniveau von  $\alpha = 0,05$  vorgeschlagen.

Problematisch bei diesem Ansatz ist jedoch die Fallzahl-Abhängigkeit der Entscheidung. Bei unter diesem Aspekt kleineren Gruppen könnte die Prozedur stoppen,

bei größeren nicht, auch wenn identische Lageunterschiede vorliegen. Ein Regressions- oder Klassifikationsbaum mit höheren Gesamt- bzw. Knotenfallzahlen bei einem Split würde somit komplexer ausfallen als ein Baum mit kleineren Anzahlen.

Grundsätzlich problematisch bei der Festlegung von Stopp-Regeln ist zudem, dass hinter dem Stopp liegende Assoziationsstrukturen unentdeckt bleiben, beispielsweise eine Wechselwirkung zwischen Variablen, die einzeln betrachtet keinen erkennbaren Beitrag zur Partitionierung liefern.

Um die beschriebenen Probleme der Stopp-Kriterien zu bewältigen, wird bereits bei Breiman [4] ein mehrstufiges Verfahren vorgeschlagen. Hier wird die Partitionierung soweit fortgesetzt, bis die Größe der Endknoten recht klein geworden ist. Anschließend kann der so gewonnene (komplexe) Baum wieder zurückgestutzt werden, indem Mindestumfänge für die endgültigen Endknoten, die Mindestunterschiede zwischen den Endknoten sowie eine minimale Reinheit innerhalb der Knoten gefordert wird. Weitergehende Kriterien, etwa die Definition von Kosten-Funktionen („cost-complexity pruning“) für Fehlentscheidungen, werden bei Breiman diskutiert.

## 2.6.4 Diskussion und Fazit

CART-Analysen zeichnen sich durch einfache Handhabung, gute Interpretierbarkeit und Flexibilität aus. Zudem lassen sich neu hinzukommende Subjekte in Bezug zur Zielgröße anhand des Baumes sofort in eine Gruppe (einen Knoten) einordnen, der etwa als Hoch- oder Niedrig-Risikogruppe bezeichnet werden könnte. Dies ist mit (generalisierten) linearen Modellen nicht möglich, da – wie im Begriff bereits angedeutet – stets lineare Zusammenhänge modelliert werden.

Andererseits geben CART-Verfahren – im Gegensatz zu linearen Modellen – keine Schätzer für die Effekte der einzelnen Faktoren an. Zudem gibt es keine im Algorithmus eingebettete Möglichkeit zur Analyse in hierarchischen Strukturen (wie etwa bei Messwiederholungen gegeben), feste oder zufällige Faktoren oder zur Prüfung auf mögliche Wechselwirkungen zwischen mehreren Faktoren. Aufgrund der Verästelungen in der Baumstruktur lassen sich für letzteren Aspekt jedoch Hinweise ableiten. In zweistufigen Verfahren können – wie bei Dahms [8] vorgeschlagen – zunächst alle Effekte fest und ohne Cluster-Struktur modelliert werden, und im zweiten Schritt Variationsstrukturen innerhalb der Endknoten (durch Anpassung eines Varianzkomponenten-Modells) untersucht werden.

Ein weiterer Nachteil der Klassifikations- und Regressionsbäume, insbesondere im Vergleich zu klassischen linearen Modellen, stellt die zwingende Dichotomisierung von nicht-binären Merkmalen dar. Insbesondere bei stetigen Einflussgrößen entsteht durch die Einteilung in zwei Gruppen ein Informationsverlust, wodurch die Modellgüte in der Regel negativ beeinflusst wird.

Da beim CART-Algorithmus immer univariate Splits betrachtet werden, gibt es keine Aussage über die globale Optimalität des berechneten Baumes. Zur Modelldiagnostik können Methoden der Kreuz-Validierungsverfahren, Residualanalyse oder „leave-out“-Ansätze benutzt werden, die bei Dahms näher beschrieben sind.

Im Grundansatz ist die CART-Methode auf Daten von kontinuierlichem oder klassifiziertem Skalenniveau beschränkt. Die Anpassung eines Regressionsbaums auf Überlebenszeiten, d.h. Zeit bis zu einem Ereignis bzw. bis zum Ende der Beobachtung („Zensierung“) ist daher nicht ohne Informationsverlust möglich. Einen Vorschlag zur Anwendung auf Überlebensdaten ist aber bei Pashova und Ulm [42] nachzulesen.

Als Fazit wird die CART-Analyse als Werkzeug aufgefasst, um einen ersten Überblick über die Datenstruktur zu erhalten. Innerhalb einer Analysestrategie können die mittels des CART-Algorithmus gefundenen relevanten Variablen (bzw. Splits) mit denjenigen verglichen werden, die mit linearen Modellen gefunden wurden. Mit den gewonnen Informationen über relevante Einflussfaktoren und den angesprochenen Hinweisen auf Wechselwirkungen kann im Anschluss daran ein geeignetes lineares Modell angepasst werden.

---

## 3 Analysestrategie

Basierend auf den in Kapitel 2 vorgestellten Methoden soll in diesem Kapitel eine Analysestrategie für Registerdaten allgemein und speziell für den Beispieldatensatz, der im folgenden Kapitel vorgestellt und ausgewertet wird, entwickelt werden.

### 3.1 Wahl der statistischen Modellklasse

Um eine Registerdatenbank oder eine ähnliche Datenquelle zur Darstellung von Behandlungsqualität und -erfolg auszuwerten, sollte nach ausführlicher Betrachtung der Erhebungsmethode und der Datenstruktur zunächst eine grobe Plausibilisierung der Daten erfolgen, falls dies nicht im Zuge der Datenerfassung bereits geschehen ist.

Zum Zwecke der Datendeskription sollten zwar einfache Darstellungen der beobachteten Mittelwerte bzw. Häufigkeiten hinsichtlich der Zielgröße erfolgen, jedoch sollten diese keinesfalls als primäre Grundlage für die Ergebnisdarstellung dienen. Vielmehr müssen die beobachteten Effekte der Kliniken bezüglich geeigneter Kovariaten – nämlich solchen, die den Behandlungserfolg beeinflussen – angepasst werden.

Wie in Kapitel 2 bereits diskutiert wurde, müssen zusätzlich zur Risikoadjustierung Auswirkungen der Varianzaufblähung und der hierarchischen Datenlage berücksichtigt werden. Dies sollte durch die Anpassung von **gemischten Modellen** und der Modellierung der Einflüsse der Kliniken als **zufällige Effekte** geschehen. Je nach dem Skalenniveau der Zielgröße führt dieser Ansatz zu klassischen (vgl. Kapitel 2.3) bzw. generalisierten gemischten linearen Modellen (vgl. Kapitel 2.4).

Da diesen Modellen zahlreiche Voraussetzungen und Annahmen zugrunde liegen, können bei Zweifeln an diesen Bedingungen zusätzlich verteilungsfreie bzw. robuste

Verfahren zur Anwendung kommen, die zwar in der Regel weniger flexibel einsetzbar sind, jedoch meist deutlich schwächere Forderungen an Modell-Voraussetzungen stellen. Hierzu kann bei Registerdaten etwa der in Kapitel 2.6 vorgestellte **CART**-Algorithmus benutzt werden.

## 3.2 Modellanpassung und Variablenselektion zur Risikoadjustierung

Da die Anpassung eines gemischten Modells ohne Hinzunahme von Kovariaten einen fairen Vergleich zwischen den Einrichtungen nicht ermöglicht, müssen in den Teil der festen Effekte  $X\beta$  neben einem Gesamtmittel  $\mu$  bzw.  $\beta_0$  diejenigen Kovariaten und mögliche Wechselwirkungen einbezogen werden, von denen ein relevanter Einfluss auf die Zielgröße bekannt ist oder zumindest vermutet wird.

Eine Ungleichheit in der Verteilung der relevanten Faktoren zwischen den Kliniken ist bei Registerdaten in nahezu allen Fällen gegeben. Die Modellanpassung bewirkt dann, dass die Einflüsse, die durch die risiko-relevanten Kovariaten erklärt werden, nicht mehr fälschlicherweise dem Zentrum zugerechnet werden.

Für die Auswahl der Kovariaten stehen im Wesentlichen die in den folgenden Abschnitten beschriebenen Konzepte zur Auswahl.

### 3.2.1 Auswahl über Datenexploration, robuste Verfahren

Wie in Kapitel 2.6.3.2 erläutert wurde, kann der CART-Algorithmus zwar kaum zur primären Analyse benutzt werden, da zufällige Effekte und hierarchische Strukturen schwer abzubilden sind. Wenn aber wenig Vorinformation über mögliche Einflussfaktoren verfügbar sind, kann dieses Instrument zur Datenexploration und zur Aufdeckung etwaiger Wechselwirkungen oder anderer komplexer Strukturen zwischen den (mit festen Effekten) modellierten Kovariaten jedoch hilfreich sein.

Alle Variablen und Kombinationen, die nach geeigneter Stützung des Regressions- oder Klassifikationsbaums im Modell verbleiben, können somit zur Analyse ins finale gemischte lineare Modell übernommen werden.

### 3.2.2 Auswahl nach Signifikanz, Selektionsverfahren

Für alle bisher beschriebenen Modellansätze lassen sich die Modelle über verschiedene Verfahren so anpassen, dass nur noch diejenigen Parameter im Modell verbleiben, die einen relevanten Beitrag zur Erklärung der Streuung des Zielparameters  $Y$  leisten. Parameter, die keinen im Sinne der statistischen Signifikanz nennenswerten Beitrag leisten, werden eliminiert bzw. nicht eingeschlossen.

Darüber hinaus muss die Modellwahl Wechselwirkungen zwischen diesen Parametern berücksichtigen. Hinweise auf diese können beispielsweise aus bivariaten Betrachtungen oder auch aus den Ergebnissen robuster Verfahren (hier CART-Analyse) abgeleitet werden.

Folgende Methoden stehen hierbei zur Verfügung:

**Volles Modell:** Diese Methode ist die Standardmethode („Full Model Fit“; keine Selektion), bei der alle in der Modellgleichung aufgeführten Parameter bzw. Wechselwirkungsterme einbezogen werden. Häufig ist dies die einzige Methode, die zur Betrachtung der Signifikanz nötig ist, beispielsweise wenn nur wenige Parameter ohne größere Interdependenzen betrachtet werden.

**Vorwärts-Selektion:** Bei der Vorwärts-Selektion („Forward Selection“) wird zu Beginn der Prozedur ein Modell ohne unabhängige Variable spezifiziert.

Für jede Einflussgröße errechnet die Vorwärtsmethode nun F-Statistiken, die den Beitrag der Variablen zum Modell angeben, wenn sie enthalten ist. Die p-Werte für diese F-Statistiken werden mit einem Eingangswert  $f_\alpha$  verglichen, der in der Modellanweisung spezifiziert wird (z.B.  $f_\alpha = 0,05$ ). Wenn der p-Wert von keiner F-Statistik die vorgegebene Signifikanzschwelle  $f_\alpha$  unterschreitet, stoppt die Selektion. Andernfalls fügt die Methode die Variable dem Modell hinzu, die die F-Statistik mit dem größten Wert besitzt.

Die Methode errechnet dann wieder F-Statistiken für die Variablen, die nicht einbezogen wurden, und der Auswahlprozess wird wiederholt. So werden die Variablen nacheinander dem Modell hinzugefügt, bis keine außerhalb des Modells verbliebene Variable eine F-Statistik besitzt, deren p-Wert unterhalb des vorgegebenen  $f_\alpha$  liegt.

Ist eine Variable einmal ins Modell einbezogen, bleibt sie enthalten.

**Rückwärts-Selektion:** Die Rückwärts-Selektion („Backward Elimination“) startet mit der Berechnung der F-Statistiken für ein Modell, welches zunächst alle unabhängigen Variablen einschließt.

Dann werden Variablen nacheinander aus dem Modell gelöscht, bis alle im Modell verbliebenen Parameter F-Statistiken zeigen, deren p-Werte unterhalb einer festgelegten Schwelle  $b_\alpha$  (z.B.  $b_\alpha = 0,1$ ) liegen. An jedem Schritt wird die Variable, die den kleinsten Beitrag zum Modell liefert, gelöscht.

**Schrittweise Selektion:** Die Schrittweise Selektion („Stepwise Selection“) stellt eine Modifikation der Vorwärts-Selektions-Technik dar und unterscheidet sich dadurch, dass Variablen, die ins Modell eingeschlossen wurden, nicht notwendigerweise dort verbleiben.

Wie in der Vorwärts-Selektion werden Variablen nacheinander dem Modell hinzugefügt, deren F-Statistik einen p-Wert  $p < f_\alpha$  zeigen. Nachdem jedoch eine Variable hinzugefügt wurde, betrachtet die schrittweise Methode alle Variablen, die bereits im Modell eingeschlossen wurden, und löscht jede Variable, deren F-Statistik einen p-Wert oberhalb einer Schwelle  $b_\alpha$  zeigt. Erst nach dieser Überprüfung und dem nötigen Entfernen von Variablen kann die nächste Variable dem Modell hinzugefügt werden.

Die schrittweise Prozedur endet,

- wenn keine der Variablen außerhalb des Modells eine F-Statistik zeigt, die verglichen mit  $f_\alpha$  signifikant ist UND jede Variable im Modell verglichen mit  $b_\alpha$  Signifikanz zeigt, ODER,
- wenn die Variable, die dem Modell hinzugefügt werden soll, erst im vorangegangenen Schritt gelöscht wurde.

Darüber hinaus stehen weitere Verfahren, wie „Maximum  $R^2$ -Verbesserung“, „Minimum  $R^2$ -Verbesserung“, „ $R^2$ -Selektion“, „adjustierte  $R^2$ -Selektion“ oder „Mallows'  $C_p$ -Selection“ zur Auswahl. Übersichtsarbeiten zur Modellselektion finden sich bei Hocking [26] sowie Judge [29].

Die dargestellten Methoden weisen hinsichtlich des Resultats Unterschiede auf. So tendiert die Vorwärts-Methode eher zur Überspezifizierung, während die Rückwärts-methode eher unterspezifiziert.

Ein Hauptkritikpunkt an der Auswahl nach Signifikanz ist die Fallzahlabhängigkeit. In der Praxis sind bei hohen Fallzahlen in der Regel viele der betrachteten

Kovariaten im Modell hinsichtlich eines  $\alpha$ -Niveaus (häufig  $\alpha = 0,05$ ) signifikant, obwohl ihr Einfluss kaum ausgeprägt ist. Andererseits können, bedingt durch zu kleine Datenumfänge oder hohe Streuungen, wichtige Effekte unentdeckt bleiben.

### 3.2.3 Auswahl nach Einflussgrad

Um die Fallzahlabhängigkeit bei der Auswahl zu umgehen, kann alternativ zur Signifikanz als Kriterium der geschätzte Koeffizient (Steigungsparameter bzw. Odds Ratio) des Einflussfaktors benutzt werden. Hierzu sollten möglichst a-priori Kriterien aufgestellt werden, die zum Einschluss in die Hauptanalyse berechtigen.

Gerade Letzteres wird in der Praxis häufig schwer realisierbar sein, da der Umgang mit und die Interpretation von Koeffizienten-Schätzern ein gewisses Maß an statistischer Vorstellungskraft einerseits und medizinisch-fachliches Wissen andererseits erfordern.

### 3.2.4 Auswahl nach medizinisch-fachlichen Gesichtspunkten, externe Informationsquellen

Die bisher in diesem Abschnitt vorgestellten Optionen zur Parameterauswahl berücksichtigen ausschließlich statistische Kriterien und sind im Rahmen einer Analyse im klinischen Bereich für sich alleine genommen nicht hinreichend. Die zu bevorzugende Methode zur Auswahl der Adjustierungsparameter stellt vielmehr die Selektion nach fachlichen Gesichtspunkten, d.h. nach bekannten medizinischen und epidemiologischen Erkenntnissen, dar. Ohne interdisziplinäre Zusammenarbeit besteht die Gefahr, dass in der Analyse wichtige substanzwissenschaftliche Aspekte unberücksichtigt bleiben.

Da bei den meisten Indikationen in Registerdaten relevante Kovariate bei den medizinischen Experten bereits bekannt sind, kann zur Variablenauswahl Beratung durch diese Personengruppe – beispielsweise durch Vertreter der teilnehmenden Einrichtungen oder auch durch von der Untersuchung unabhängige Experten – in Anspruch genommen werden. Ist es nicht möglich, eine solche Informationsquelle zu

erschließen, können die gewünschten Informationen alternativ auch über Literaturrecherchen erfolgen, falls zu der im Register betrachteten Indikation genügend Literatur vorliegt.

Ein weiterer Vorteil der Informationsgewinnung durch externe Quellen liegt darin, dass so die Vorauswahl und die Analyse voneinander getrennt werden können und nicht zufällige oder singuläre Zusammenhänge eine zu hohe Bedeutung erhalten.

### 3.3 Test-Prozedur

In der vorliegenden Aufgabenstellung interessiert man sich zunächst für die Frage, ob die (wie oben beschrieben geeignet adjustierten) Ergebnisse zwischen den Kliniken Unterschiede aufweisen. Diese Fragestellung, d.h. die Prüfung der globalen Hypothese  $H_0 : \sigma_a^2 = 0$  im gemischten linearen Modell, kann mittels geeigneter Z-Statistiken, deren kritische Grenzen erforderlichenfalls mittels Simulationen ermittelt werden können (siehe auch Kapitel 2.4.2.3), zu einem Signifikanzniveau von üblicherweise  $\alpha = 0,05$  beantwortet werden.

In einem nächsten Schritt sind klinik-spezifische Effekte zu prüfen. Hierbei kann durchaus diskutiert werden, ob *überhaupt* einzelne Kliniken auf Unterschiede gegen den (geschätzten) Mittelwert geprüft werden sollten. Möchte man beurteilen, ob tatsächlich einzelne Kliniken im Vergleich zum Gesamtmittel besonders gute oder besonders schlechte Behandlungsqualität liefern, sind Signifikanztests über individuelle Klinikeffekte jedoch unerlässlich. Als Grundlage dafür, dass Abschätzungen für den wahren Effekt einer Klinik vorgenommen werden können, ist davon auszugehen, dass ein beobachtetes Patientenkollektiv eines Klinikums als Stichprobe aus allen möglichen Patienten interpretiert werden. Vor dem Hintergrund, dass die am Register teilnehmenden Einrichtungen ebenfalls als Stichprobe aus allen möglichen Kliniken aufgefasst werden können, erscheint diese Annahme gerechtfertigt. Im Rahmen der hier vorgelegten Analysestrategie wird somit vorgeschlagen, Vertrauensbereiche und Tests für die Effekte einzelner Einrichtungen anzugeben.

Lässt sich die im ersten Absatz dieses Abschnitts gestellte Frage nach generellen Unterschieden zwischen Kliniken nicht bejahen, wird im Zuge der Analyse jedoch

nicht mehr geprüft, welche individuellen Klinik-Effekte signifikant von 0 verschieden sind. Eine Darstellung der adjustierten EBLUP-Schätzer mit Konfidenzintervallen sollte dennoch vorgenommen werden. Falls jedoch auf Unterschiede zu einem vordefinierten Signifikanzniveau (z.B.  $\alpha = 5\%$ ) geschlossen werden kann, möchte man wissen, *welche* Kliniken – als Faktoren mit zufälligen Effekten modelliert – sich signifikant vom Durchschnitt unterscheiden. Es werden also Vergleiche zwischen jeweils einer Kliniken mit allen anderen Kliniken angestellt. Da die weitestgehende Fragestellung bereits bejaht wurde, d.h. beim Globaltest wurde die Nullhypothese  $H_0$ : „Es gibt keine Unterschiede zwischen den Zentren“ verworfen, kann auf der nächsten Stufe wieder zum selben Niveau  $\alpha$  getestet werden. Auf eine weitergehende Multiplen-Korrektur sollte hier verzichtet werden, um die Güte der – ohnehin recht konservativen – Klinik-spezifischen Tests nicht weiter dramatisch zu verringern.

Trotzdem kann der Fall eintreten, dass zwar  $H_0$  verworfen wurde, jedoch keine einzelne Einrichtung Signifikanz zeigt. In diesem Fall muss davon ausgegangen werden, dass generell unterschiedliche Behandlungserfolge zwischen Kliniken im allgemeinen vorliegen, jedoch von den teilnehmenden Kliniken keine nachweislich „schlechter“ oder „besser“ als der Durchschnitt behandelt.

### 3.4 Umgang mit fehlenden und unplausiblen Werten

In Registerdatenbanken sind in vielen Fällen Kovariaten nur lückenhaft erfasst. Subjekte, die in wenigstens einer der ins Modell einbezogenen Variablen keine vollständige Erfassung aufweisen, müssen vielfach aus der Analyse ausgeschlossen werden, da viele Modelle Vollständigkeit voraussetzen. Teilweise ist es jedoch möglich, die Ausprägung „fehlend“ als zulässige Werte mit zu betrachten und somit die Zahl der auswertbaren Subjekte zu erhöhen.

Das Fehlen von Werten kann vielfältige Ursachen und Auswirkungen besitzen. Können Ausfälle als zufällig angenommen werden (diese Frage muss im Einzelfall untersucht und abgewogen werden), bewirken die fehlenden Werte lediglich eine Verringerung der Testgüte und keine Verzerrung der Ergebnisse. Im Falle des Beispieldatensatzes dieser Arbeit, wenn etwa die Zielgröße das Überleben des Patienten darstellt, können jedoch Abhängigkeiten zwischen dem Erfassungsstatus der Kovariaten und der Zielgröße selbst bestehen. In diesen Fällen muss damit gerechnet werden, dass die Ergebnisse nicht mehr unverfälscht darstellbar sind.

Auch können fehlerhafte Einträge eine mitunter große Auswirkung auf das Ergebnis besitzen, insbesondere wenn im Modell Normalverteilung vorausgesetzt wird (d.h., es werden Mittelwerte betrachtet) und fehlerhafte Angaben Ausreißer in den Daten erzeugen und somit Mittelwerte stark verändern.

Im Vorfeld der Auswertung sollte somit in Zusammenarbeit mit den Trägern der Registerdatenbank dafür gesorgt werden, dass die Anzahl der fehlenden und falschen Eingaben so gering wie möglich ist, um die Zahl der auszuschließenden Subjekte und die Verfälschung durch Datenfehler zu minimieren. In jedem Fall sollte vor der Analyse eine – zumindest grobe – Plausibilitätsprüfung erfolgen.

Die Problematik der fehlenden Werte ist in Zusammenhang mit der Variablen-selektion zu sehen. Grundsätzlich können bei großen Datenbeständen durchaus viele Kovariaten in das Analysemodell einbezogen werden, da die Gefahr eines „overfits“ dann gering ist. Andererseits erhöht die Zahl der Kovariaten bei lückenhaften Datenlagen auch die Zahl der nicht mehr einschließbaren Subjekte. Ist dies der Fall, sollte versucht werden, mit der Auswahl der Variablen möglichst sparsam vorzugehen. Eine Kombination aus schrittweiser Selektion und Auswahl nach fachlichen Gesichtspunkten ist hier ein möglicher Ansatz.

Tritt eine Situation ein, in der lückenhaft erfasste Einflussgrößen von den Experten als unverzichtbar für die Risikoadjustierung angesehen werden, kann zunächst versucht werden, die Daten durch Nachfrage bei den teilnehmenden Einrichtungen nachträglich zu beschaffen. Falls hierdurch keine befriedigende Datenvollständigkeit erzielt werden kann, müsste mit Datenersetzungsregeln („data imputation“) versucht werden, die Datenvollständigkeit zu gewährleisten. Auf dem Feld der Datenersetzung wurden zahlreiche Methoden und Optionen vorgeschlagen, was die Auswahl der Regeln einerseits recht aufwändig gestalten kann. Zum anderen wird in die Regeldefinitionen zumeist ein Anteil an Subjektivität einfließen, da Ersetzungsregeln stets auf die jeweilige medizinisch-fachliche Situation angepasst werden müssen.

## 3.5 Darstellung der Ergebnisse

Im Methodenteil dieser Arbeit wurde bereits vielfach darauf hingewiesen, dass die Darstellung der rohen, d.h. unadjustierten, Mittelwerte bzw. Ereignishäufigkeiten

keinesfalls als Hauptanalyse veröffentlicht werden sollte. Zwar ist bei der Datenaufbereitung zunächst eine Betrachtung der wichtigsten Kennzahlen für die teilnehmenden Einrichtungen darzustellen, diese sollte jedoch ausschließlich deskriptiven Charakter besitzen. Erst das Zusammenspiel aus Risikoprofilen der Kliniken und Kennzahlen hinsichtlich der Zielgröße ist als valide Darstellung zu sehen. Bei der Analyse sollte also, erst nachdem die Variablenselektion erfolgte, die Darstellung der Klinik-spezifischen Risikoprofile und der Ergebnisse des Hauptanalyse-Modells erfolgen.

Bei der Aufbereitung der Ergebnisse gibt es mehrere Darstellungsmöglichkeiten, von denen im Rahmen dieser Arbeit nur einige benutzt werden sollen. So werden die Ergebnisse des Hauptanalyse-Modells nach Prüfung der Globalhypothese zunächst in tabellarischer Form mit adjustiertem Klinikeffekt, entsprechendem Konfidenzintervall und Signifikanzniveau dargestellt. Bei generalisierten Modellen – d.h. wenn der Klinikeffekt nicht auf der Originalskala geschätzt wurde – sollten die Ergebnisse auf das Skalenniveau der Originalwerte transformiert werden, um bessere Interpretierbarkeit zu gewährleisten.

Um eine Kombination von Risikoprofilen und Ergebnissen zu erreichen, können die mittels  $X\hat{\beta}$  im marginalen Modell geschätzten Populationskenngrößen je Einrichtung dargestellt werden. Erwartete Mittelwerte (oder Raten) und beobachtete Mittelwerte (oder Raten) werden dann gegenübergestellt.

### 3.6 Anmerkungen zur Analyse-Software

Zur Analyse von gemischten Modellen existieren vielfältige Software-Lösungen. Die meisten dieser Programmpakete werden von kommerziellen Anbietern bereitgestellt und ständig weiterentwickelt. Aufgrund der Komplexität und Vielfalt der verfügbaren Schätzalgorithmen, Optimierungsverfahren und weiteren Optionen ist es häufig nicht möglich, mit zwei unterschiedlichen Software-Paketen für einen identischen Datensatz identische Lösungen zu erhalten. Dieser Umstand kann zwar einerseits nicht befriedigen, andererseits aber unterscheiden sich die Ergebnisse oft nur marginal und führen selten zu gänzlich unterschiedlichen Interpretationen.

In dieser Arbeit wurden die Simulationen und Analysen der gemischten Modelle mit der Software SAS<sup>®</sup> durchgeführt. Je nach Skalenniveau der Zielgröße und Modellklasse werden hierbei unterschiedliche Prozeduren aufgerufen. Verwendet wurden

PROC MIXED (gemischte lineare Modelle) und PROC GLIMMIX (generalisierte gemischte lineare Modelle). Darüber hinaus existieren viele weitere Software-Pakete, die beispielsweise auf der Internet-Seite des „Centre for Multilevel Modelling“ der Universität Bristol einzeln vorgestellt und diskutiert werden [66].

Bei der Anpassung des CART-Algorithmus auf die Daten wurde auf die nicht kommerzielle Software **R**, die dem kommerziellen Programm S-PLUS weitgehend entspricht, zurückgegriffen.

---

## 4 Anwendungsbeispiel: Das Berliner Herzinfarktregister

In diesem Kapitel werden Ergebnisse der in Kapitel 2 beschriebenen Methodologie unter Einbeziehung des in Kapitel 3 skizzierten Analysekonzepts vorgestellt. Zur Anwendung kamen dabei die Daten aus der Beispieldatenbank des Berliner Herzinfarktregisters (BHiR).

Bei der Vorstellung der Ergebnisse wird zunächst auf die wichtigsten medizinischen Grundlagen der zugrunde liegenden Population eingegangen. Der Kern der Darstellung wird jedoch die Diskussion der statistischen Methoden hinsichtlich ihrer praktischen Anwendbarkeit, Eigenschaften und Ergebnisse sein.

### 4.1 Medizinische Grundlagen

In diesem Abschnitt werden medizinische und epidemiologische Grundlagen für das Anwendungsbeispiel dieser Arbeit, dem akuten Myokardinfarkt, dargestellt. Es wird auf die Häufigkeit der Erkrankung und die damit verbundene sozioökonomische Bedeutung, auf die wichtigsten Risikofaktoren und die hauptsächlich betroffenen Bevölkerungsgruppen eingegangen.

Als Herzinfarkt wird der „Untergang“ (das Absterben) von Herzmuskelgewebe nach einem plötzlich auftretenden kompletten Verschluss einer oder mehrerer Koronararterien bezeichnet.

Gelegentlich kann auch eine nicht totale Stenose, d.h. eine nur unvollständige Verengung einer Koronararterie, zum Infarkt führen. In der überwiegenden Mehrzahl der Fälle liegt dem Koronargefäßverschluss eine höhergradige Stenose mit Ausbildung einer Thrombose zugrunde, wobei die Thrombose in der Regel durch die

Ruptur arteriosklerotischer Gefäßablagerungen verursacht wird. Man unterscheidet einen Vorderwand-, Hinterwand- und Seitenwandinfarkt, einen Scheidewandinfarkt (Septuminfarkt) sowie Kombinationsinfarkte. Die Infarkte mit der schlechtesten Prognose sind Vorderwand- und Scheidewandinfarkte. Wird ein drohender Infarkt festgestellt, muss so schnell wie möglich – und zwar innerhalb weniger Stunden – eine Therapie erfolgen, andernfalls entstehen irreparable Schäden in der mit sauerstoffreichem Blut unterversorgten Herzregion.

### 4.1.1 Anatomie und Physiologie

Das menschliche Herz wird von der linken und rechten Herzkranzarterie versorgt, die als Koronararterien bezeichnet werden. Die linke Herzkranzarterie (= Arteria coronaria sinistra) versorgt den vorderen Bereich des Herzens, die rechte (= Arteria coronaria dextra) den hinteren. Die linke Herzkranzarterie verzweigt sich wenige Zentimeter nach ihrem Abgang aus der Aorta in zwei Äste, den so genannten Ramus circumflexus. Dieser durchblutet die linke Herzkammer sowie den Ramus interventricularis anterior, der das Septum und, wenn auch in geringerem Umfang, die linke Herzkammer versorgt.

Die Herzkranzarterien bilden, wie alle Arterien, Abzweigungen und Verästelungen bis hin zu den Kapillaren, welche durch den Herzmuskel ziehen und diesen mit Nährstoffen und Sauerstoff versorgen. Außerdem helfen sie, CO<sub>2</sub> und andere Stoffwechselprodukte abzutransportieren.

Aufgrund von arteriosklerotischen Ablagerungen an den Wänden der Arterien – laienhaft als „Verkalkung“ bezeichnet – reicht der Blutdurchfluss oft nur noch unzureichend dafür aus, den Herzmuskel mit Blut und mit Sauerstoff zu versorgen. Sollte sich das Gefäß völlig verschließen, ist, wie erwähnt, ein Herzinfarkt die Folge. Die Verkalkung ist ein komplexer Prozess, bei dem eine Reihe verschiedener – auch erblich bedingter – Faktoren eine Rolle spielt.

Nach dem Gesetz von Hagen-Poiseuille, das für laminare (d.h. nicht turbulente) Strömungen gilt, verringert sich der Blutfluss mit der vierten Potenz in Abhängigkeit vom Gefäßradius. Wird ein starres Gefäß z.B. um 50% verengt, so sinkt der Blutfluss daher auf ein Sechzehntel. Da die Koronararterien nicht starr sind und in ihnen auch keine laminare Strömung herrscht, gilt dieses Gesetz nur als grobe Annäherung an die wahren Verhältnisse, zeigt aber die Dramatik starker Verengungen.

**Abbildung 4.1: Darstellung der linken Herzkammer in der Kernspintomographie nach einem Herzinfarkt**



#### 4.1.2 Epidemiologie und Burden of Disease

In den Industrieländern machen die Herz- und Gefäßkrankheiten (ICD-10 Katalog, Kapitel IX) zusammen etwa 40 bis 50 Prozent aller Todesursachen aus und stellen damit nach wie vor die größte Gruppe, noch deutlich vor den bösartigen Neubildungen (Krebserkrankungen). Laut einer Pressemitteilung des Statistischen Bundesamts vom 14. Februar 2005 fielen im Jahre 2003 etwa 396.000 der insgesamt 853.000 Todesfälle in Deutschland auf eine Kreislauferkrankung als Todesursache [72].

Die Sterblichkeitsrate (Letalität) des akuten Herzinfarktes (ICD-10, Kapitel IX, Ischämische Herzkrankheiten I20 - I25) sinkt zwar in den letzten Jahrzehnten in einigen europäischen Ländern und den USA. Die Bedeutung ist jedoch nach wie vor groß. So entsprach die Todesursache bei rund 20% der Frauen und Männer im Jahre 1997 den Folgen einer koronaren Herzkrankheit.

In Deutschland erleiden jährlich rund 300 Personen je 100.000 Einwohner, also etwa 250.000 Menschen, einen Herzinfarkt. Die Letalität wird für die westlichen Industrienationen gegenwärtig auf etwa 25 bis 30 Prozent geschätzt, was einer Sterbezahl von 60.000 bis 75.000 in Deutschland entspricht. Im Jahre 2003 verstarben laut Todesursachenstatistik 64.229 Personen (davon 64% Männer) am akuten Myokardinfarkt [73]. Im Krankenhaus allerdings versterben nur noch zwischen 4% und 12% der Patienten.

Bei der Herzinfarkt-Letalität wird häufig eine Grenze von 28 Tagen als entscheidend für die Prognose angesehen, da davon ausgegangen wird, dass ein Patient, der innerhalb dieser Zeitspanne nicht verstirbt, das Ereignis selbst zunächst einmal überlebt hat. Ein wichtiger Faktor für die Herzinfarkt-Letalität stellt die Zeitspanne vor dem Erreichen der Klinik („Prähospitalzeit“) dar: Etwa ein Drittel der Infarkt-Patienten versterben innerhalb dieser Zeitspanne. Aber auch die Patienten, die nach einer zu langen Zeitspanne das Notfallklinikum lebend erreichen, haben schlechtere Chancen, ihren Infarkt zu überleben.

Allerdings muss bei der Betrachtung der Letalität berücksichtigt werden, dass viele Patienten aufgrund eines früheren Infarkts eine Herzinsuffizienz oder eine andere Folgeerkrankung oder Komplikation entwickeln und später daran, und nicht primär am Infarkt selbst, versterben. Schätzungen über die Gesamt-Sterblichkeit belaufen sich auf zwischen 90.000 und 200.000 Personen jährlich.

Bei den Kosten für die Gesellschaft muss zwischen direkten, also den medizinisch abgerechneten Krankheits- und Behandlungskosten, und den indirekten (bzw. volkswirtschaftlichen) Kosten unterschieden werden. Tabelle 4.1 zeigt anhand der entstehenden jährlichen Behandlungs- und Folgekosten die Bedeutung der Herz-Kreislauf-Erkrankungen für das deutsche Gesundheitssystem sowie den Einfluss des Alters der Patienten auf die Kosten [74].

Somit entfällt etwa ein Sechstel (16,2% bzw. 35 Mrd. €) aller Ausgaben im Gesundheitswesen auf Herz-Kreislauf-Erkrankungen. Ein Fünftel dieser Kosten (etwa 7 Mrd. €) entstehen durch ischämische Herzkrankheiten. Beim akuten und rezidivierenden (d.h. wiederholten) Infarkt werden insgesamt 1,26 Milliarden Euro ausgewiesen [70].

Vergleicht man bei den Herz-Kreislauf-Erkrankungen die Gruppe der über 65-jährigen mit den unter 65-jährigen, so stellt man fest, dass die Kosten pro Einwohner in der älteren Gruppe nahezu 10-mal so hoch ist wie in der jüngeren Gruppe. Dieses Verhältnis ist größer als bei jeder anderen Erkrankungsart (Gesamtverhältnis alt/jung: 3,7). Die über 65-jährigen stellen zur Zeit etwa ein Sechstel der Gesamtbevölkerung dar, verursachen aber zwei Drittel der Behandlungskosten. Legt man den demographischen Trend mit zunehmendem Anteil der älteren Bevölkerung zugrunde, ist von weiter stark steigenden Gesamtkosten auszugehen.

Die Kosten, die der Volkswirtschaft durch Arbeitsunfähigkeit, Invalidität und Tod entstehen, sind schwer zu beziffern. Da sich ein großer Teil der direkten Kosten auf

**Tabelle 4.1: Kosten 2004 nach Krankheitsklassen und Alter in € je Einwohner der jeweiligen Altersgruppe**

Gegenstand der Nachweisung		Gesamt	Altersgruppe [in Jahren]			
			< 15	15-65	65-85	≥ 85
<b>Krankheiten insgesamt</b>		2.730	1.110	1.980	5.950	14.750
I	Infektiöse und parasitäre Krankheiten	50	70	40	70	110
II	Neubildungen	210	20	140	620	860
III	Krankheiten d. Blutes u. d. blutbildenden Organe	10	10	10	30	60
IV	Endokrine, Ernährungs- und Stoffwechselkrankh.	140	20	100	400	450
V	Psychische und Verhaltensstörungen	280	120	210	450	2.720
VI	Krankheiten des Nervensystems	120	30	80	290	730
VII	Krankheiten d. Auges u. d. Augenanhangsgebilde	70	100	40	150	250
VIII	Krankheiten d. Ohres u. d. Warzenfortsatzes	30	40	20	60	90
<b>IX</b>	<b>Krankheiten des Kreislaufsystems</b>	430	10	190	1.430	3.690
X	Krankheiten des Atmungssystems	140	190	100	230	370
XI	Krankheiten des Verdauungssystems	400	80	410	630	600
XII	Krankheiten der Haut und der Unterhaut	40	40	40	70	120
XIII	Krankheiten des Muskel-Skelett-Systems	300	30	240	690	1.050
XIV	Krankheiten des Urogenitalsystems	100	20	90	220	240
XV	Schwangerschaft, Geburt und Wochenbett	40	0	60	–	–
XVI	Zustände mit Ursprung in der Perinatalperiode	10	70	0	0	0
XVII	angeborene Fehlbildungen, Deformitäten	10	40	10	10	10
XVIII	Symptome und klinisch abnorme Befunde a.n.k.	130	70	50	240	2.330
XIX	Verletzungen und Vergiftungen	130	60	90	270	940
XXI	Faktoren, die den Gesundheitszustand beeinflussen	70	80	60	100	130

die nicht mehr berufstätige Bevölkerungsgruppe verteilt, kann von vergleichsweise geringen indirekten Kosten ausgegangen werden.

Einen wichtigen Ansatzpunkt bei der Kostenvermeidung stellt sicherlich die Kontrolle der allgemein bekannten Risikofaktoren wie Übergewicht, Hypertonie, Nikotin- und Alkoholabusus sowie Stress dar. Verbesserte – und damit teurere – Behandlungsmöglichkeiten stehen zwar nicht der Prävention, jedoch der Reduktion von Folgekosten entgegen.

Bei der Therapie des akuten Myokardinfarkts ist zum einen die weitere Senkung der Letalität in der Prähospitalphase von Bedeutung. Andererseits kann seitens der Kliniken, deren Behandlungsqualität hinsichtlich des Therapieerfolgs, gemessen bspw. als Letalität *nach* der Einlieferung des Patienten, bestimmt werden soll, auf die Prähospitalphase kaum Einfluss genommen werden. Zu dieser Phase ist auch die Zeitspanne zwischen dem Auftreten des Infarkts und erster Notfalltherapie bzw. die

Zeitspanne zwischen Auftreten des Infarkts und Eintreffen in der Klinik zu zählen. Daher kann im Rahmen dieser Arbeit beim Vergleich zwischen den Notfall-Kliniken auf die Behandlung *vor* der Einlieferung nur im Sinne einer Kovariaten bzw. einer Risikoadjustierung eingegangen werden.

## 4.2 Das Berliner Herzinfarktregister – Vorbetrachtungen

### 4.2.1 Beschreibung des Patientenkollektivs

Das Berliner Herzinfarkt Register (BHiR) existiert seit 1996. Seit Januar 1999 sind kontinuierlich Daten von mehreren Tausend Herzinfarkt-Patienten in verschiedenen teilnehmenden Kliniken gesammelt und in einer zentralen Datenbank abgelegt worden.

Seit dem Beginn der Aufzeichnungen wurde der zugrunde liegende Fragebogen mehrmals geändert, was zu unvollständiger Erfassung bestimmter Variablen führte. Für die Datenanalyse im Rahmen dieser Arbeit wurden vom BHiR Daten aus dem Zeitraum vom 1. Januar 1999 bis zum 31.12.2003 zur Verfügung gestellt, innerhalb dessen insgesamt 6.972 Patienten erfasst wurden. Innerhalb dieses Zeitraums gab es eine Änderung des Fragebogens.

Die grundsätzliche Zielstellung des Registers ist die Betrachtung von initial nach dem Infarkt aufgenommenen und im Zentrum behandelten Patienten hinsichtlich ihres Überlebens im Krankenhaus selbst als primäre Zielgröße. Aber auch andere (sekundäre) Zielgrößen werden immer wieder diskutiert.

Um Betrachtungen über die Zeitachse für alle Kliniken berücksichtigen zu können, wurden diejenigen Kliniken entfernt, bei denen nicht mehr als 20 Fälle pro Jahr dokumentiert wurden und bei denen keine kontinuierliche Erfassung über den gesamten vierjährigen Zeitraum erfolgten. Es verblieben somit Daten von 10 Kliniken und 4.055 gemeldeten Fälle in der Datenbank. Von diesen Fällen wurden noch Patienten aufgrund von

- intrahospitalen Infarkten,

- Hinzuverlegung aus einem anderen Krankenhaus,
- schnelle Weiterverlegung (z.B. in das Deutsche Herzzentrum Berlin [DHZB]) innerhalb von 48 Stunden, ohne dass Angaben über das Überleben oder die weitere Therapie verfügbar waren, oder
- fehlenden Angaben hinsichtlich des Überlebens (ja/nein) oder der demographischen Kenngrößen Alter und Geschlecht

aus der Auswertung ausgeschlossen (siehe folgende Tabelle).

**Tabelle 4.2: Patienten-Stichprobe des BHiR**

	absolut	%
1. Fälle insgesamt (Jan 1999 - Dez 2003)	6.972	100,0
2. Ausgeschlossene Fälle nach Klinikum		
- Klinikum hat zu wenige Fälle pro Jahr berichtet	1.026	14,7
- Klinikum hat nicht im gesamten Zeitraum teilgenommen	1.891	27,1
Summe:	2.917	41,8
3. Bereinigte Stichprobe	4.055	100,0
4. Ausgeschlossene Fälle nach Patient		
- Patient erlitt intrahospitalen Infarkt	234	5,8
- Patient wurde hinzuverlegt (keine Ersteinweisung)	180	4,4
- Patient wurde weiterverlegt (innerhalb von 48 Stunden)*	79	1,9
- Patienten-Angaben zu Tod oder Demographie fehlen	97	2,4
Summe:	590	14,5
5. Realisierte Stichprobe	3.465	85,5

\* und keine Angaben über weitere Therapie bzw. Tod verfügbar

## 4.2.2 Charakteristika der Datenbank

### 4.2.2.1 Erfasste Parameter

Die Patientendaten wurden im Betrachtungszeitraum auf einem 4-seitigen Fragebogen im Klinikum dokumentiert (die verwendeten Bögen können auf der Internetseite des BHiR eingesehen werden [64]), vor Ort geprüft (monitoriert) und nach einer Qualitäts- bzw. Plausibilitätsprüfung in einer gemeinsamen Datenbank erfasst.

Wie oben bereits angemerkt, wurde der Fragebogen innerhalb des Betrachtungszeitraums verändert, wodurch eine Transformation einiger nicht durchgehend auf die gleiche Weise erhobenen Variablen nötig wurde (siehe folgende Übersicht).

In der folgenden Übersicht sind die erfassten Variablen auf dem Fragebogen des BHiR thematisch dargestellt. Den vollständigen Bogen kann man auf der Internetseite des BHiR abrufen.

0. Angaben zur Klinik (Klinikcode, Patientenummer)
1. Patientendaten:
  - Demographische Angaben (Alter, Körpermaße (Gewicht, Größe), Geschlecht, Wohnort (PLZ), Staatsangehörigkeit, Familienstand, berufliche Stellung (bei der Erhebung),
  - Angaben zum Infarktbeginn,
  - Angaben zur Verlegung des Patienten,
  - Klinikankunft des Patienten.
2. Präexistierende Risikofaktoren (z.B. arterielle Hypertonie, Diabetes Mellitus, Rauchen);
3. Präexistierende Erkrankungen (z.B. bekannte koronare Herzkrankheit [KHK], früherer Herzinfarkt);
4. Akutdiagnostik;
5. Akuttherapie;
6. Komplikationen (z.B. Reinfarkt oder Tod);

7. Diagnostik und Therapie;
8. Intrahospitale Komplikationen;
9. Entlassung/Verlegung;
10. 30-Tages-Letalität;
11. 1-Jahres-Letalität.

Die – durchaus zahlreichen – Änderungen am Erhebungsbogen sind zum Teil auf veränderte medizinische Erkenntnisse, zum anderen aber auch auf Erfahrungen, die während der Erhebung gewonnen wurden, zurückzuführen. Für eine Auswertung an einem konsolidierten Register, insbesondere bei Betrachtungen über die Zeitachse, ist es natürlich wünschenswert, dass keine Änderungen in der Erhebung durchgeführt werden, da andernfalls bestimmte Informationen nicht berücksichtigt werden können, was sich auf die Aussagekraft der Ergebnisse negativ auswirken kann.

#### 4.2.2.2 Plausibilitätsprüfung

In durch Spenden und Sponsorengelder geförderten öffentlichen bzw. teilprivaten Registern ist es aufgrund der hohen Zahl der Dateneinträge und der begrenzten Kapazitäten bei der Dateneingabe und -aufbereitung nicht immer möglich, alle Angaben der Kliniken auf Plausibilität zu prüfen. Zu den Fehlern in den Quelldaten kommen meist Eingabefehler hinzu, die häufig gar nicht oder erst bei der Analyse auffallen. Manche unauffällige Fehler können auch gänzlich unentdeckt bleiben. Auch bei großen Datenmengen können bereits einzelne Fehleingaben zu verfälschten Ergebnissen führen; insbesondere bei kontinuierlichen Variablen verfälschen durch Eingabe- oder Messfehler verursachte Ausreißer zuweilen gravierend das Ergebnis.

Daher wurde im Rahmen der Auswertung in dieser Arbeit darauf geachtet, Fehleingaben zu identifizieren und – falls möglich – zu korrigieren bzw. – falls nicht möglich – zu eliminieren. Beispielsweise wurde eine 66-jährige Patientin mit einem Body Mass Index (BMI), der als

$$\text{BMI [kg/m}^2] = \frac{\text{Körpergewicht [kg]}}{\text{Körpergröße [m]}^2}$$

definiert ist, von  $82,5 \text{ kg/m}^2$  erfasst (Körpergröße 1,05m; Gewicht 91kg). Da es sich hier offensichtlich um einen Eingabefehler handelt (vermutlich bei der Körpergröße), wurde der BMI für dieses Subjekt eliminiert.

#### 4.2.2.3 Fehlende Werte

Bei parametrischen Modellanpassungen (klassische bzw. gemischte generalisierte lineare Modelle) werden häufig vollständige Patientendaten vorausgesetzt. In der vorliegenden Datenbank des BHiR ist jedoch – außer für die Parameter Alter, Geschlecht sowie der Zielgröße – eine teils recht hohe Anzahl von fehlenden Werten zu beobachten, was die Anzahl zu betrachtender Fälle je nach modellierten Einflussgrößen stark reduziert.

Weiterhin stellt sich die Frage, ob die Tatsache, dass Werte fehlen, als ein rein *zufälliger* Effekt aufgefasst werden kann, oder ob nicht Werte bei bestimmten Patienten oder auch bei bestimmten Zentren *systematisch* fehlen. Bei zufälligen oder „systemneutralen“ Ausfällen können die gewonnenen Schätzungen dann als unverfälscht gelten, wenn auch mit reduzierter Präzision.

Neben einer Vielzahl von anderen Methoden, sind im Umgang mit fehlenden Werten die folgenden Alternativen vorstellbar:

- fehlende Werte werden für die Auswertung nicht berücksichtigt;
- „fehlend“ wird – zumindest bei kategoriellen Variablen – als zulässige Ausprägung aufgefasst;
- fehlende Werte werden mittels Imputationsmechanismen (Datenersetzungsregeln) ersetzt.

In dieser Arbeit wird die Analyse primär basierend auf der ersten und zweiten Alternative aus der obigen Aufzählung durchgeführt. Die Problematik der nicht zufälligen Ausfälle wird an den relevanten Stellen im einzelnen diskutiert.

## 4.3 Deskriptive Statistik und nicht adjustierte Analysen

Vorab sei bemerkt, dass die in dieser Arbeit dargestellten Kliniknummern aus der Datenbank zum Zwecke der Verblindung zunächst zufällig angeordnet und anschließend von 1 bis 10 neu vergeben wurden.

Aufgrund der Vielzahl von erfassten Merkmalen, der teils hohen Rate fehlender Werte und wegen der teils nicht einheitlichen Erfassung über die beiden Versionen des Erhebungsbogens hinweg konnten für die Auswertung nicht alle Variablen betrachtet werden.

Die hinsichtlich der Versorgungs- bzw. Behandlungsqualität interessierende Hauptzielgröße „Krankenhaus-Letalität“ ist als binäre Variable, d.h. mit den Ausprägungen *Patient ist im Krankenhaus nach Einlieferung mit akutem Herzinfarkt*

- *verstorben* bzw.
- *nicht verstorben*.

erfasst worden.

Von den 3.465 dokumentierten Patienten (s.o.) sind insgesamt 397 Patienten (oder 11,46%) im Krankenhaus verstorben. Eine erste „naive“ Darstellung zeigt die Sterberaten für die 10 teilnehmenden Krankenhäuser (Tabelle 4.3).

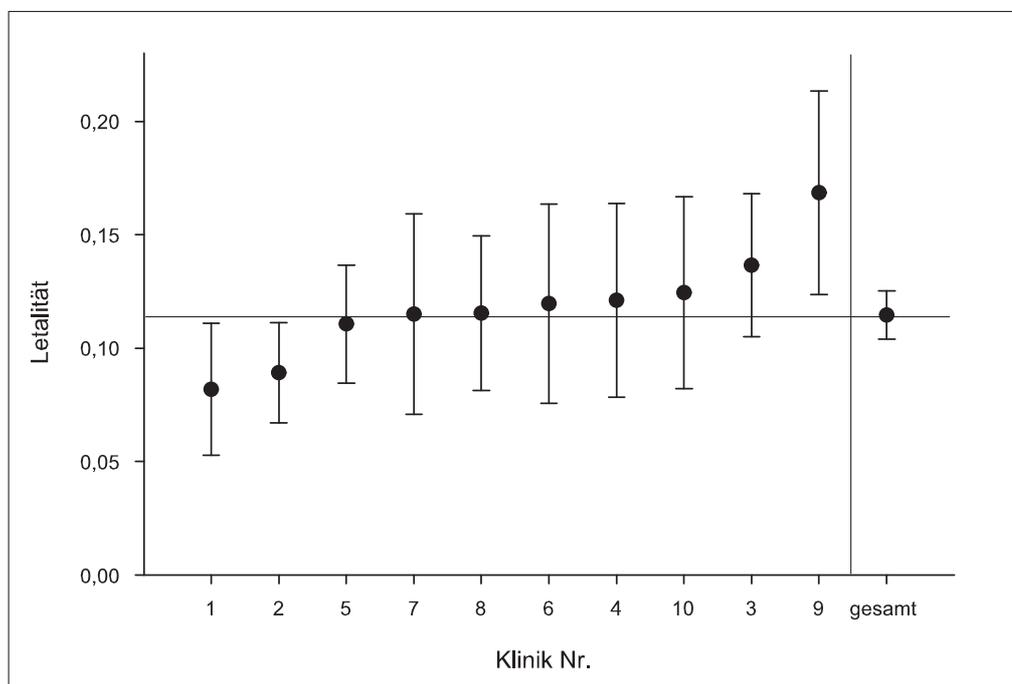
Obwohl im Rahmen dieser Arbeit der Faktor „Zeit (Jahr)“ nicht untersucht wird, sei an dieser Stelle der Verlauf der beobachteten Letalitätsraten über die Jahre im Betrachtungszeitraum zumindest erwähnt (siehe Tabelle 4.4).  $n$  bezeichnet hier die Zahl der im jeweils Jahr eingeschlossenen Fälle des betreffenden Klinikums. Da einige Kliniken nicht über alle Jahre des Beobachtungszeitraums teilnahmen bzw. die Einschlusskriterien für die Analyse verletzten, kann in einigen Fällen keine Letalitätsrate angegeben werden.

Während sich insgesamt – d.h. über alle Kliniken hinweg – im Vergleich der Jahre bei leicht abnehmender Tendenz recht ähnliche Sterberaten zeigen, schwanken die Ergebnisse innerhalb der Zentren von Jahr zu Jahr zum Teil stark. Beispielsweise

**Tabelle 4.3: Beobachtete Krankenhaus-Letalität nach Klinik**

Klinik Nr.	Anzahl Fälle insgesamt	Letalität		95%-Konfidenzgr.*	
		absolut	relativ	untere	obere
1	342	28	8,19 %	5,28%	11,09%
2	639	57	8,92 %	6,71%	11,13%
3	454	62	13,66 %	10,50%	16,82%
4	223	27	12,11 %	7,83%	16,39%
5	560	62	11,07 %	8,47%	13,67%
6	209	25	11,96 %	7,56%	16,36%
7	200	23	11,50 %	7,08%	15,92%
8	338	39	11,54 %	8,13%	14,94%
9	267	45	16,85 %	12,36%	21,34%
10	233	29	12,45 %	8,21%	16,69%
Gesamt	3.465	397	11,46 %	10,40%	12,52%

\* Schätzfehler wurde separat, d.h. unadjustiert, bestimmt

**Abbildung 4.2: Krankenhaus-Letalität nach Klinik – separate 95%-Vertrauensintervalle**

**Tabelle 4.4: Krankenhaus-Letalität nach Klinikum und Jahr**

Klinik #	1999		2000		2001		2002		2003	
	n	Letal.								
1	55	5,5%	62	12,9%	77	9,1%	94	7,4%	54	5,6%
2	155	11,6%	126	4,8%	137	13,9%	123	7,3%	98	5,1%
3	102	11,8%	102	15,7%	86	15,1%	91	15,4%	73	9,6%
4	69	7,2%	54	16,7%	40	12,5%	36	16,7%	24	8,3%
5	101	12,9%	111	6,3%	120	9,2%	126	14,3%	102	12,7%
6	0	-, -	58	19,0%	67	7,5%	57	12,3%	27	7,4%
7	9	22,2%	64	12,5%	59	13,6%	40	5,0%	28	10,7%
8	41	14,6%	101	11,9%	113	10,6%	83	10,8%	0	-, -
9	61	9,8%	77	23,4%	43	9,3%	53	18,9%	33	21,2%
10	60	11,7%	79	13,9%	36	8,3%	58	13,8%	0	-, -
Gesamt	653	11,0%	834	12,7%	778	11,2%	761	11,8%	439	9,6%

zeigt sich bei Klinik Nr. 9 in den Jahren 1999 und 2001 eine im Vergleich zur Gesamtrate etwas geringere Letalität, während sie in den übrigen Jahren deutlich höhere Sterberaten aufweist. Andere Kliniken wiederum – wie etwa Klinikum Nr. 8 – veränderten sich bezüglich ihrer Letalität weniger stark.

### 4.3.1 Separate Betrachtung der Kliniken

Bei der Betrachtung über alle Jahre hinweg (Tabelle 4.3) fällt auf, dass das Krankenhaus mit der höchsten Letalität (Klinik 9) im Vergleich zu der Klinik mit der geringsten Letalität (Klinik 1) eine etwa doppelt so hohe Rate aufweist (relatives Risiko = 2,1). Es ist zu untersuchen, ob die beobachteten Schwankungen zwischen den Kliniken dem Zufall geschuldet sein können oder auf tatsächliche Unterschiede zurückzuführen sind.

Wie in Abbildung 4.2 dargestellt ist, liegen zwei Kliniken (Nr. 1 und 2) unter Berücksichtigung der Stichprobengröße mit den Letalitätsraten unterhalb des Durchschnitts; eine Klinik (Nr. 9) liegt oberhalb. Bei den übrigen Kliniken können keine signifikanten Unterschiede im Vergleich zum Durchschnitt festgestellt werden.

### 4.3.2 Nicht adjustierte Modellierung nach Klinikum im GLM

Um der Tatsache Rechnung zu tragen, dass die Patienten nicht  $p = 10$  **verschiedenen** Populationen, sondern gemeinsam **einer** Population entstammen, werden alle Patienten und Kliniken in **einem** Modell betrachtet. Zu diesem Zweck wird wegen der dichotomen Eigenschaft der Zielgröße ein klassisches logistisches Regressionsmodell (siehe auch Kapitel 2.2.4) mit folgender Modellbeziehung angepasst:

$$P(Y = 1) = \frac{1}{1 + e^{-X\beta}} \quad \text{mit} \quad X\beta = \beta_0 + \sum_{i=1}^k \beta_i X_i ,$$

wobei  $Y \in \{0, 1\}$  die möglichen Ausprägungen

- $Y = 0$ : Das Ereignis tritt nicht ein:  
„der Patient verstirbt im Krankenhaus nicht“ und
- $Y = 1$ : Das Ereignis tritt ein:  
„Der Patient verstirbt im Krankenhaus“

annahmen kann.

In einer ersten Betrachtung wird als Einflussgröße nur das Klinikum als Faktor mit festen Effekten berücksichtigt; es wird also zunächst ein logistisches Modell mit nur einer Einflussgröße  $X_1$  angepasst:

$$X\beta = \beta_0 + \beta_1 X_1, \quad \text{mit } X_1: \text{Klinikum in } p = 10 \text{ Stufen .}$$

Da die Einflussgröße „Klinik“ mit nominalem Skalenniveau vorliegt, kann hier von einer „logistischen ANOVA“ anstatt einer logistischen Regression gesprochen werden.

Zur Prüfung, ob tatsächlich Unterschiede zwischen den 10 Kliniken bestehen, wird ein Score-Test im logistischen Modell für den Hypothesensatz

$$H_0 : \beta_1 = 0 \quad \text{vs.} \quad H_1 : \beta_1 \neq 0$$

zum Niveau  $\alpha = 0,05$  durchgeführt. Dieser entspricht dem Chi-Quadrat-Test mit  $p - 1 = 9$  Freiheitsgraden auf Unterschiede in den Letalitätsraten zwischen den Kliniken, also

$$\begin{aligned} H_0 : & \quad p_i = p_j && \text{für alle } i \neq j \in (1, \dots, p) && \text{vs.} \\ H_1 : & \quad p_i \neq p_j && \text{für wenigstens ein } i \neq j \in (1, \dots, p), \end{aligned}$$

mit  $p_i = P(Y = 1 | X_1 = i)$ : wahre unbekannte Ereigniswahrscheinlichkeit im Klinikum  $i$ .

Der Score-Test (entspricht dem  $\chi^2$ -Test in der 2x10-Tafel) mit  $p - 1 = 9$  Freiheitsgraden zeigt mit einer empirischen Signifikanz von  $p = 0,0358$  Unterschiede zwischen den Kliniken, die o.g. Nullhypothese wird also zum 5%-Niveau verworfen.

Die Parameterschätzung für  $\beta_0$  (Absolutglied/Intercept: „Grundsterblichkeit“ im linearen Prediktor) und  $\beta_1$  ( $X_{1i}$ : Effekt der Klinik) – zerlegt in 10 Designvariablen – für dieses Modell sind in Tabelle 4.5 dargestellt.

**Tabelle 4.5: Krankenhaus-Letalität nach Klinikum – nicht adjustierte Modellierung im GLM**

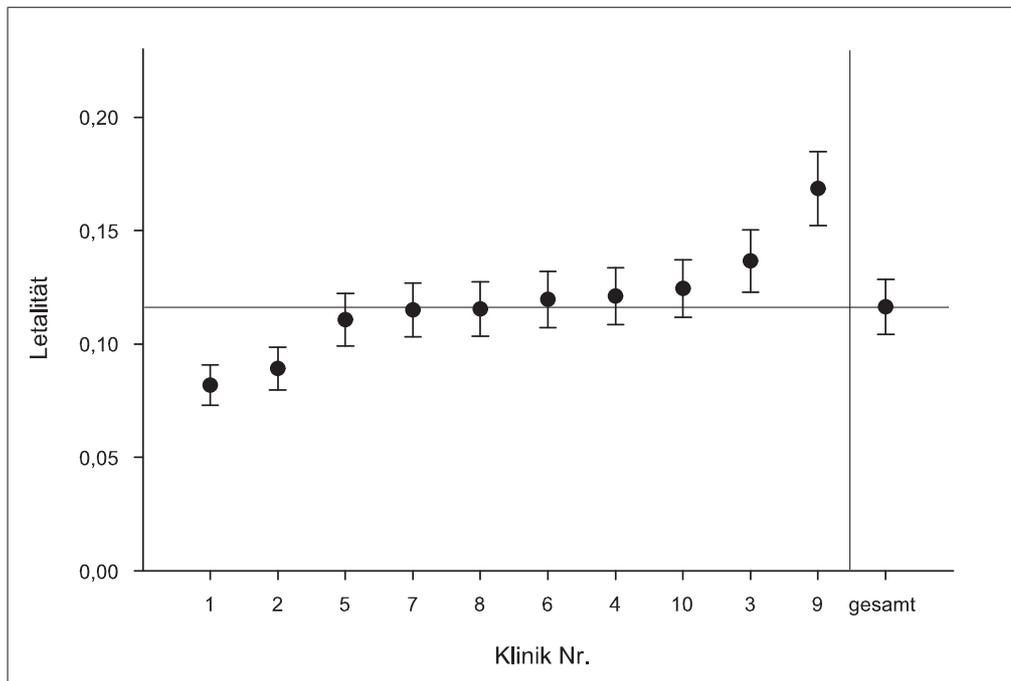
Parameter	Koeffizient			OR Schätzer	Letalität			
	Schätzer	UG*	OG*		Schätzer	UG*	OG*	
$\hat{\beta}_0$ (Intercept)	-2,0271	-2,1391	-1,9151	—	0,1164	0,1054	0,1284	
$\hat{\beta}_{1i}$ (Klinik)	1	-0,3901	-0,7535	-0,0266	0,6770	0,0819	0,0738	0,0907
	2	-0,2963	-0,5642	-0,0285	0,7436	0,0892	0,0805	0,0987
	3	0,1830	-0,0815	0,4475	1,2008	0,1366	0,1239	0,1503
	4	0,0448	-0,3321	0,4217	1,0458	0,1211	0,1097	0,1335
	5	-0,0564	-0,3177	0,2050	0,9452	0,1107	0,1002	0,1222
	6	0,0310	-0,3591	0,4211	1,0315	0,1196	0,1083	0,1319
	7	-0,0136	-0,4180	0,3908	0,9865	0,1150	0,1041	0,1269
	8	-0,0098	-0,3286	0,3090	0,9903	0,1154	0,1044	0,1273
	9	0,4311	0,1234	0,7388	1,5389	0,1685	0,1534	0,1848
	10	0,0763	-0,2892	0,4418	1,0793	0,1245	0,1128	0,1372

\* untere und obere Grenzen des 95% Konfidenzintervalls nach Wald

Die Odds Ratios der  $p$  verschiedenen Stufen von  $X_1$

$$\widehat{\text{OR}}_i = e^{\hat{\beta}_{1i}} \quad (i = 1, \dots, p)$$

**Abbildung 4.3: Krankenhaus-Letalität nach Klinikum**  
– GLM-basierte 95%-Vertrauensintervalle



stellen hier den Vergleich zum Durchschnitt dar. Eine Klinik mit einer exakt durchschnittlichen beobachteten Letalitätsrate würde einen Effekt von  $\hat{\beta}_{1i} = 0$  und demnach ein geschätztes Odds Ratio von 1 zeigen.

Im Vergleich zur separaten Betrachtung bleiben die Punktschätzer der Letalität natürlich unverändert; jedoch sind die Vertrauensintervalle aufgrund der modellbasierten Varianzschätzung deutlich kleiner. Dies führt dazu, dass nun die geschätzte Letalitätswahrscheinlichkeit auch von Klinik Nr. 3 signifikant über dem Durchschnitt liegt. Nach einem ersten naiven Ranking würden also die Kliniken Nr. 1 und 2 als – verglichen mit dem Durchschnitt – im Sinne des Überlebens der Patienten besonders erfolgreich, und die Kliniken 3 und 9 als weniger erfolgreich klassifiziert werden. Für die übrigen sechs Kliniken würden keine Unterschiede zur Gesamtwahrscheinlichkeit nachgewiesen werden können.

Paarweise Vergleiche (Odds Ratios) zwischen zwei beliebigen Kliniken  $i$  und  $j$

können nach den Ergebnissen aus Tabelle 4.5 nun gemäß der Beziehung

$$\begin{aligned}\widehat{\text{OR}}_{(ij)} &= \frac{\hat{p}_i/(1-\hat{p}_i)}{\hat{p}_j/(1-\hat{p}_j)} = \frac{\widehat{\text{OR}}_i}{\widehat{\text{OR}}_j} = \frac{e^{\hat{\beta}_{1i}}}{e^{\hat{\beta}_{1j}}} \\ &= e^{\hat{\beta}_{1i}-\hat{\beta}_{1j}}\end{aligned}$$

geschätzt werden.

Diese Darstellung kann jedoch aus vielen Gründen nicht befriedigen, unter anderem deshalb, weil nicht berücksichtigt wurde, inwieweit dieser Vergleich „fair“ ist; es fehlt also die Risikoadjustierung der beobachteten Letalitätsraten nach dem individuellen Risiko, welches die Patienten tragen. Zudem ist die hierarchische Struktur und die zufällige Eigenschaft der Zentrumseffekte hier noch nicht berücksichtigt. Da diese einfache Darstellung – d.h. ohne Berücksichtigung von Kovariaten – wie beschrieben die Gegebenheiten der Praxis nicht hinreichend abbildet, sollten die beobachteten Ergebnisse nur als Einführung verstanden werden.

### 4.3.3 Übergang zum Generalisierten Gemischten Linearen Modell

Die vorigen unadjustierten Darstellungen sind lediglich als Voranalyse und Deskription der Daten zu verstehen, da bisher noch keine Anpassung auf die hierarchische Situation und keine Berücksichtigung der Zufallsauswahl der Kliniken erfolgte.

Als Modellklasse der Wahl wird – wie in Kapitel 2 bzw. 3.1 besprochen wurde – ein Gemischtes Generalisiertes Lineares Modell (GLMM) benutzt. Mittels dieser Anpassung wurden EBLUP-Schätzer mit 95%-Konfidenzintervallen für die Klinikeffekte  $\gamma_i$  bestimmt. Eine erste unadjustierte Modellierung mit gewählter Logit-Linkfunktion hat somit die folgende Gestalt (siehe auch Kapitel 2.4.2.2):

$$\text{logit}(p) = X\beta + Z\gamma + e = \beta_0 + \gamma_i + e .$$

Die Anpassung, durchgeführt mit der Software PROC GLIMMIX (eine Anpassung der Konvergenzkriterien war hier nicht notwendig), zeigt das im Folgenden dargestellte Ergebnis.

Zunächst wurde gemäß des Analysekonzepts die Nullhypothese  $H_0 : \sigma_a^2 = 0$  mittels des selbst konstruierten Signifikanztests (vgl. Kapitel 2.4.2.3) zum Niveau  $\alpha = 0,05$  geprüft. Für die Z-Statistik wurde der folgende Wert ermittelt:

$$Z = \frac{\hat{\sigma}_a^2}{\text{SE}(\hat{\sigma}_a^2)} = \frac{0,0285}{0,0261} = 1,0897 . \quad (4.1)$$

Da  $Z > c_\alpha = 0,9596$  ist, wird  $H_0$  hier zum Niveau  $\alpha = 0,05$  verworfen. Mittels der empirischen Verteilungsfunktion für Z unter  $H_0$ , gilt für den p-Wert  $p \approx 0,023$ . In diesem **nicht adjustierten** Modell kann nun durch Prüfung der 10 Einzelhypothesen

$$H_{0i} : \gamma_i = 0 \quad \text{vs.} \quad H_{1i} : \gamma_i \neq 0$$

geprüft werden, ob und welche Klinik-Effekte signifikant von 0 verschieden sind. Die Ergebnisse sind in Tabelle 4.6 dargestellt.

**Tabelle 4.6: Krankenhaus-Letalität nach Klinikum – nicht adjustierte Modellierung im GLMM**

Parameter	$n$	Koeffizient im LP			Letalität (Originalskala)			p-Wert	
		Schätzer	$k_u^*$	$k_o^*$	Schätzer	$k_u^\#$	$k_o^\#$		
$\hat{\beta}_0$ (Intercept) <sup>^</sup>		-2,0297	-2,1804	-1,8791	0,1161	0,1015	0,1325	—	
$\hat{\gamma}_i$ (Klinik)	1	342	-0,1723	-0,4241	0,0794	0,0996	0,0792	0,1245	0,1796
	2	639	-0,1785	-0,4031	0,0460	0,0990	0,0807	0,1209	0,1191
	3	454	0,1109	-0,1188	0,3405	0,1280	0,1045	0,1559	0,3439
	4	223	0,0190	-0,2446	0,2826	0,1181	0,0933	0,1484	0,8876
	5	560	-0,0329	-0,2582	0,1923	0,1128	0,0921	0,1374	0,7744
	6	209	0,0129	-0,2536	0,2794	0,1174	0,0925	0,1480	0,9243
	7	200	-0,0040	-0,2728	0,2648	0,1157	0,0909	0,1462	0,9766
	8	338	-0,0036	-0,2500	0,2429	0,1158	0,0928	0,1435	0,9775
	9	267	0,2158	-0,0339	0,4654	0,1402	0,1127	0,1730	0,0902
	10	233	0,0328	-0,2286	0,2941	0,1195	0,0946	0,1499	0,8058

\* untere und obere Grenzen des  $i$ -ten 95% Konfidenzintervalls

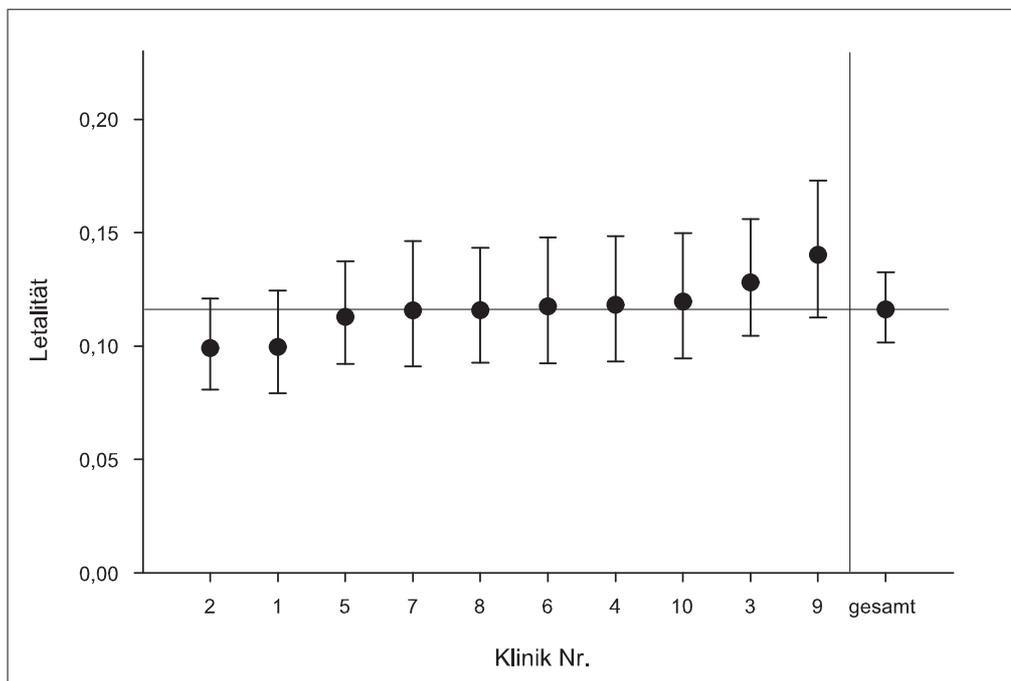
# mittels inverser Link-Transformation berechnet

<sup>^</sup> im marginalen Modell

Es zeigt sich, dass trotz der Ablehnung der globalen Nullhypothese – d.h. es wird auf Unterschiede zwischen Kliniken geschlossen – keine in die Stichprobe gelangte

Klinik als signifikant „besser“ oder „schlechter“ als die geschätzte Grundwahrscheinlichkeit eingestuft werden konnte. Klinik Nr. 9 zeigte hier den höchsten (schlechtesten) und Klinik Nr. 2 den kleinsten (besten) nicht adjustierten Schätzwert für  $\gamma_i$  bzw. für  $p_i$ . Jedoch zeigen beide zum zweiseitigen Testniveau von 5% keine Signifikanz. Somit ist das Ergebnis des ersten GLMM fundamental verschieden von denen des klassischen logistischen Regressionsmodells (vgl. Kapitel 4.3.2).

**Abbildung 4.4: Krankenhaus-Letalität nach Klinikum**  
– GLMM-basierte 95%-Vertrauensintervalle



Im Vergleich zum klassischen logistischen Regressionsmodell bleiben beim GLMM zwar die Vorzeichen der Schätzer (im linearen Prediktor), da  $\hat{\sigma}_a^2 > 0$  ist, für jedes Klinikum erhalten. Jedoch verringerte sich durch den Schrumpfungseffekt der EBLUP-Schätzung die Variabilität der Punktschätzer. Weiterhin „tauschen“ Klinik Nr. 1 und 2 den Rangplatz: Das im GLM „beste Klinikum“ (Klinik Nr. 1) weist nach der Schrumpfung einen höheren Schätzwert als Klinikum Nr. 2 auf, da Letzteres aufgrund der deutlich höheren Fallzahl geringer zur Gesamtwahrscheinlichkeit bewegt wurde.

**Bemerkung**

Beim Übergang auf die Originalskala (dargestellt in den letzten drei Spalten von Tabelle 4.6) wurden die Schätzer und Konfidenzgrenzen aus dem linearen Prediktor entsprechend der Logitfunktion transformiert. Die Konfidenzgrenzen der transformierten Ereigniswahrscheinlichkeiten zeigen **nicht** die Konfidenzgrenzen für die prognostizierte Ereigniswahrscheinlichkeit der Kliniken, da hier die Variabilität von  $X\hat{\beta}$  nicht berücksichtigt wurde. Die entsprechenden Intervalle lassen sich nur für das Modell ohne Kovariaten darstellen, da andernfalls für jedes Individuum unterschiedliche Ausprägungen in  $X\hat{\beta}$  vorliegen. Im vorliegenden Fall der einfachen logistischen Regression sind die Unterschiede zwischen diesen beiden Typen von Konfidenzintervallen allerdings sehr gering.

## 4.4 Auswahl von Kovariaten zur Risikoadjustierung

Nach dieser ersten Voranalyse der Krankenhaus-Letalität muss gemäß des Analysekonzepts im nächsten Schritt geklärt werden, um welche relevanten Kovariaten das Auswertungsmodell erweitert werden muss, um für die teilnehmenden Kliniken einen möglichst fairen Vergleich zu gewährleisten. Hierbei muss demnach entschieden werden, welche der in Abschnitt 3.2 vorgestellten Methoden verwendet werden soll.

Zum Zeitpunkt des Beginns der Datenanalyse für diese Arbeit wurde zunächst keine medizinisch-fachliche Information zur Risikoadjustierung verwendet. Es lag zwar umfangreiches Literaturmaterial zu Risikofaktoren bei der Krankenhaussterblichkeit nach Myokardinfarkt (siehe beispielsweise bei Maier [37] [38] oder Theres [54]) bzw. nach anderen Eingriffen (z.B. bei Ugolini zum Bypass [55]) vor, jedoch war seitens des BHiR noch keine Empfehlung zur Variablenauswahl verfügbar. Daher wurde zunächst mit der Auswahl nach Signifikanz bzw. Auswahl durch Datenexploration, dies in einem ersten Schritt für die demographischen Variablen, begonnen. Im November 2006 wurde dann für die Jahre 2004 und 2005 in Zusammenarbeit mit der Universität Hamburg eine Analyse der Daten zum Zwecke des Einrichtungsvergleichs durchgeführt [65]. Für diese Analyse wurde durch Repräsentanten des Registers eine Auswahl der für die Risikoadjustierung zu verwendenden Variablen getroffen (vgl. Abschnitt 3.2.4).

Für die Endanalyse dieser Arbeit wurde diese Auswahl dann übernommen, soweit dies für die weiter zurückliegenden Jahre möglich war (siehe Abschnitt 4.4.2 ff.). Die bis dahin gewonnenen Ergebnisse zur Auswahl nach Signifikanz sind im folgenden Unterkapitel 4.4.1 daher als Exkurs dargestellt. Hier werden die vielfältigen Probleme, die durch fehlende Werte entstehen, sowie mögliche Lösungsansätze bei der Auswahl der Kovariaten diskutiert. Die Auswahl der Modellparameter erfolgt letztlich jedoch durch fachliche Beurteilung.

#### **4.4.1 Auswahl nach Signifikanz / Datenexploration für demographische Variablen (Exkurs)**

Man kann nicht direkt prüfen, wie die Zentren in dem Einrichtungsvergleich abschneiden würden, wenn alle die gleiche Population behandelt hätten. Jedoch kann unter bestimmten Voraussetzungen durch geeignete Risikoadjustierung eine Vergleichbarkeit der Zentren erreicht werden. Hierzu bietet sich an, zunächst in Erfahrung zu bringen, welche Patienten-bezogenen Kovariaten Einfluss auf die Sterblichkeit im Krankenhaus (nach einem Herzinfarkt) besitzen. Die Letalitätsraten der Zentren können dann hinsichtlich dieser Parameter adjustiert dargestellt werden.

Wichtig bei der Auswahl der Kovariaten ist es, dass sie keinen Bezug zur behandelnden Klinik besitzen, d.h. sie sollten nicht durch die Klinik selbst variierbar sein. Da es sich bei der vorliegenden Datenbank jedoch nicht um einen randomisierten Versuchsaufbau handelt, kann im allgemeinen nicht ausgeschlossen werden, dass Kliniken selbst auf die Patientenauswahl – etwa durch Ablehnung bestimmter Risikopatienten – Einfluss nehmen. Durch geeignete Adjustierung der Letalitätsraten hinsichtlich Kovariaten, die das Letalitätsrisiko beeinflussen, soll dieser Effekt, sei er durch bewusste Auswahl oder durch andere systematische oder zufällige Einflüsse verursacht, ausgeglichen werden.

Die folgenden demographischen Parameter werden in diesem Analyseschritt betrachtet (siehe auch Darstellung in 4.2.2.1; Patientendaten):

1. Alter;
2. Geschlecht;

3. Body Mass Index (BMI);
4. Staatsangehörigkeit;
5. berufliche Stellung;
6. Wohnort;
7. Familienstand.

#### 4.4.1.1 Vorbemerkungen zur Modellbildung

Für die nachfolgende Auswertung müssen die beiden Variablen berufliche Stellung und Wohnort aber zunächst näher betrachtet werden.

##### 1. Berufliche Stellung

Da die berufliche Stellung sehr stark mit dem Alter des Patienten in Zusammenhang steht (insbesondere falls „Rentner“ als eigene Ausprägung besteht), geht von der Einbeziehung dieser Variablen als solche wenig Information aus. Vielmehr sollte die berufliche Stellung den Ausbildungsstand bzw. die soziale Stellung des Patienten wiedergeben. Die Änderung des Fragebogens im Jahr 2001 trägt diesem Anspruch Rechnung, indem bei Rentnern gefragt wurde, welche Stellung zuletzt, d.h. vor der Berentung, bestand (Arbeiter, Angestellter, Beamter, Selbständiger oder Sonstiges). Die berufliche Stellung wurde dann folgendermaßen definiert:

- Falls der Infarktbeginn des Patienten vor dem Jahr 2001 lag (alter Bogen) und falls die berufliche Stellung als „Rentner“ angegeben wurde, wurde die berufliche Stellung wie *zuletzt ausgeübt*, als „nicht bekannt“ bewertet.
- Falls der Infarktbeginn des Patienten nach dem Jahr 2000 lag (neuer Bogen) und falls die berufliche Stellung als „Rentner“ angegeben wurde, wurde die berufliche Stellung wie *zuletzt ausgeübt*, abgefragt und für die Auswertung verwendet.

Da etwa die Hälfte der eingeschlossenen Fälle auf dem alten Fragebogen beruht und weil durch die Neudefinition des beruflichen Status – zusätzlich zu den ohnehin schon fehlenden Angaben – die zuletzt eingenommene berufliche Stellung aller Rentner als unbekannt und damit fehlend klassifiziert wurden, ist die Zahl der fehlenden Werte in dieser Variablen mit 1.868 (oder 54% der Fälle) sehr hoch und zudem nur für den Zeitraum nach ab dem 01.01.2001 verfügbar. Da die durchgeführte logistische Regression nur auf vollständigen Fällen berechnet wird, würde sich unter Hinzunahme des beruflichen Status ins Modell die Zahl der auswertbaren Fälle auf nur 1.240 der ursprünglichen 3.465 reduzieren. Um diesen hohen Datenverlust zu vermeiden, wurde diese Variable ausschließlich einzeln bzw. bivariat betrachtet.

## 2. Wohnort des Patienten

Der Wohnort der Patienten wurde über die Postleitzahl erfasst und durch das BHiR in die folgenden vier Gruppen eingeteilt:

1. Berlin (Ost),
2. Berlin (West),
3. PLZ > 14199,
4. PLZ < 10115.

Da eine Unterscheidung zwischen dem Postleitzahl-Bereich 3 (Hannover) und 4 (Düsseldorf) schwer interpretierbar ist, wurde für die Patienten mit Wohnort außerhalb Berlins eine neue Gruppierung vorgenommen:

3. Postleitzahlen aus Brandenburg (Berliner Umland) [69],
4. Postleitzahlen außerhalb Brandenburgs.

### 4.4.1.2 Logistische Regressionsanalyse

Um nicht zu viele Beobachtungen aufgrund von fehlenden Werten zu verlieren, wurden die genannten Parametergruppen blockweise modelliert und auf mögliche Wechselwirkungen hin untersucht.

Im logistischen Modell wird stets  $P(Y = 1)$ , also die Wahrscheinlichkeit für das Versterben des Patienten, betrachtet.

### Ergebnisse für Modell 1 (Ausschluss fehlender Werte):

Im folgenden werden die p-Werte für alle Einflussgrößen mittels Typ-3-Fehlern als feste Effekte modelliert und geschätzt. Somit hat die Reihenfolge der Parameter im Modell keinen Einfluss auf die empirische Signifikanz.

In dieses erste Modell wurden sieben Parameter (wie oben beschrieben) ohne Wechselwirkungen einbezogen, wobei alle Variablen mit Ausnahme des Alter und des BMI nominalskaliert eingingen. Die p-Werte wurden mit allen verfügbaren Werten („einzeln(1)“) und zum Vergleich basierend auf den 2.354 vollständig erfassten Fällen des Gesamtmodells („einzeln(2)“) bestimmt.

Die Ergebnisse der Modellierung sind in Tabelle 4.7 dargestellt.

**Tabelle 4.7: Logistische Regression 1 – Einfluss von soziodemographischen Variablen auf die Krankenhaus-Letalität, fehlende Werte eliminiert**

Parameter	Anzahl Werte	Werte fehlend	p-Wert Ges.-Modell	p-Wert	
				einzeln(1)	einzeln(2)
Alter [Jahre]	3.465	0	<0,0001	<0,0001	<0,0001
Geschlecht (männlich / weiblich)	3.465	0	0,6291	<0,0001	<0,0001
Body Mass Index [ $kg/m^2$ ]	2.837	628	0,7649	0,0105	0,0055
Nationalität (deutsch / nicht deutsch)	3.364	101	0,2531	0,0244	0,0242
Wohnort (Bln.-Ost/Bln.-West/Brb/sonst.)	3.267	198	0,1087	0,1042	0,0800
Familienstand (nicht allein / allein lebend)	2.970	495	0,1850	<0,0001	<0,0001
Gesamt-Modell	2.354	1.111	<0,0001	,	,
letzte berufliche Stellung (Arb./Ang./...)*	1.597	1.868	,	,	<0,0001

(1) basierend auf allen vollständigen Fällen (n=2.354)

(2) basierend auf allen verfügbaren Fällen

\* nur einzeln getestet, da Anzahl fehlender Werte aufgrund Bogens 1 zu hoch

In das Gesamt-Modell wurden alle 2.354 Fälle aufgenommen, für die alle sechs soziodemographischen Variablen erhoben worden sind. Von den insgesamt 397 *Todesfällen* (vgl. Tabelle 4.3) in der Datenbank konnten nur 214 Fälle (bzw. 54%) in diesem Modell betrachtet werden. Bei allen anderen Verstorbenen fehlten Angaben zu mindestens einem dieser Parameter.

Unter Berücksichtigung der Zahl der einschließbaren *Patienten*, wodurch noch 2.354 der 3.465 Patienten (entspricht einem Anteil von 68%) verfügbar sind, ist der Informationsverlust bei der Todesrate hoch. Die Letalitätsrate ist – anders ausgedrückt – im Modell nicht 11,5% (wie in der Gesamt-Stichprobe), sondern lediglich 9,1%. Es ist zu vermuten, dass bestimmte demographische Daten gerade aufgrund des Todes der Patienten vor dem Verlassen der Klinik nicht mehr oder nur noch mittels Befragung anderer Personen, d.h. mit größerem Aufwand erfasst werden konnten bzw. daher seltener verfügbar sind als bei den Patienten, die das Krankenhaus lebend verließen.

Vergleicht man die beiden Gruppen der vollständigen und der nicht vollständigen Fälle hinsichtlich der Letalität in einer 2x2-Tafel miteinander, so ergibt sich ein starker Zusammenhang zwischen den Merkmalen Sterblichkeit (ja/nein) und Vollständigkeit (ja/nein) mit  $p < 0,0001$  (Chi-Quadrat Test; siehe Tabelle 4.8).

**Tabelle 4.8: Zusammenhang zwischen Vollständigkeit und Mortalität**

Vollständigkeit		verstorben		nicht verstorben		Summe	
		n	↓ %	n	↓ %	n	↓ %
ja	n	214	53,9%	2.140	69,8%	2.354	67,9%
	↖ %	9,1%		90,9%		100,0%	
nein	n	183	46,1%	928	30,2%	1.111	32,1%
	↖ %	16,5%		83,5%		100,0%	
Summe	n	397	100,0%	3.068	100,0%	3.465	100,0%
	↖ %	11,5%		88,5%		100,0%	

Die Pfeile ↓ und ↖ an den Prozentzeichen bezeichnet spalten- bzw. zeilenweise Prozentuierung

Aufgrund dieser Ergebnisse kann nicht von Vorliegen „neutraler“ bzw. zufälliger Ausfälle ausgegangen werden. Vielmehr muss „informatives“ bzw. „systematisches“ Fehlen von Werten angenommen werden. Inwiefern sich die fehlenden Informationen auf die Ergebnisse hinsichtlich der primären Fragestellung (Vergleich zwischen den Kliniken) auswirken, hängt davon ab, ob dieser Ausfall-Effekt gleichmäßig über die Kliniken verteilt ist. In diesem Fall sind zwar die Ergebnisse insgesamt nicht vollständig valide, jedoch wäre zumindest ein Vergleich zwischen den Kliniken, basierend auf der vorliegenden Datenlage, zulässig.

**Ergebnisse für Modell 2 (fehlende Werte einbezogen):**

Eine Alternative, mittels derer die Zahl der fehlenden Beobachtungen im Modell reduziert werden kann, stellt ein Ansatz dar, bei der die fehlenden Werte als „erlaubte“ Ausprägung mit modelliert werden. Hierbei ist jedoch Folgendes zu beachten:

- Die Ausprägung „fehlend“ ist nur als ungeordnete Ausprägung verwendbar. Bei kontinuierlichen bzw. ordinal skalierten Variablen müsste dem fehlenden Wert eine (geordnete) Zahl zugeordnet werden, um dem Skalenniveau zu entsprechen. Im allgemeinen ist dies aber nicht möglich, da sich keine echte Zahl für die fehlenden Werte direkt anbietet; lediglich in selteneren Fällen kann beispielsweise eine Null oder ein Populationsmittelwert angenommen werden. In jedem Fall führt das Ersetzen fehlender Werte bei geordneten Variablen zu einer – potenziell strittigen – Änderung der Verteilung.
- Bei nominal skalierten oder binären und damit ungeordneten Variablen (dies sind im vorliegenden Modell die meisten) können die fehlenden Werte als Gruppe aufgefasst werden, wodurch jeder Variablen, die fehlende Werte enthält, eine zusätzliche Ausprägung zugeordnet wird. An der Verteilung innerhalb der geplanten Stufen ändert sich zunächst nichts, lediglich an den relativen Häufigkeit der Stufen.

Trotz der einfachen Implementierbarkeit können sich Schwierigkeiten bei der Interpretation ergeben. So könnte der Fall eintreten, dass sich die Beurteilungen eines Parameters bezüglich der Signifikanz (z.B. Kriterium  $p < 0,05$ ) im Vergleich zwischen beiden Modellansätzen (mit und ohne fehlende Werte) unterscheiden. Insbesondere bei einem hohen Anteil von fehlenden Werten ist dieser Fall denkbar, was wiederum auf das Vorliegen systematischer Ausfälle schließen lässt.

Um die Zahl der „missings“ zu verringern, werden jetzt im oben dargestellten logistischen Regressionsmodell aus Tabelle 4.7 für alle nominal skalierten Variablen die fehlenden Werte als zulässige Ausprägung (Stufe) berücksichtigt. Die berufliche Stellung kann somit ebenfalls in das Gesamtmodell einbezogen werden.

Zwei Parameter, das Alter und der Body Mass Index, sind von stetigem Skalenniveau, weshalb hier keine Ersetzung der fehlenden Werte vorgenommen werden kann.

Beim Alter liegen keine fehlenden Werte vor, beim BMI jedoch fehlen die Angaben bei 628 (18,1%) der insgesamt 3.465 Patienten.

Es werden zwei alternative Möglichkeiten zur Modellierung betrachtet:

1. Body Mass Index verbleibt im Modell, wodurch die verfügbare Stichprobe die um 628 Patienten verringerte Menge der 3.465 Patienten darstellt (siehe Tabelle 4.9);
2. Der BMI wird – wie auch die berufliche Stellung im Modell 1 (Tabelle 4.7) – einzeln betrachtet, wodurch ins Hauptmodell alle 3.465 Patienten einbezogen werden können (Tabelle 4.10).

**Tabelle 4.9: Logistische Regression 2 – soziodemographische Variablen mit fehlenden Werten (alle Parameter)**

Parameter	Anzahl Werte betrachtet	Anzahl (davon) fehlender Werte	p-Wert im Gesamtmodell
Alter [Jahre]	3.465	0	<0,0001
Geschlecht (männlich / weiblich)	3.465	0	0,6121
Body Mass Index [ $kg/m^2$ ]*	2.837	628	0,7345
Nationalität (deutsch / nicht deutsch)	3.465	(101)	0,3599
Wohnort (Bln.-Ost/Bln.-West/Brb/sonst.)	3.465	(198)	0,2253
Familienstand (nicht allein / allein lebend)	3.465	(495)	0,1403
letzte berufliche Stellung (Arb./Ang./...)	3.465	(1.868)	0,1347
Gesamt-Modell**	2.837	628	<0,0001

\* fehlende Werte nicht ersetzt

\*\* 628 Beobachtungen entfallen daher

Im zweiten Modell (Tabelle 4.9) konnten nun zwar schon 81,9% des gesamten Patientenkollektivs einbezogen werden, jedoch ist der Anteil der einbezogenen Verstorbenen mit  $n = 272$  (oder 68,5%) der insgesamt 397 Todesfälle noch deutlich geringer. Die in diesem Modell berücksichtigte Letalitätsrate ist somit mit 9,5% nur geringfügig höher als im ersten Modellansatz (9,1%). Die Sterblichkeitsrate ist bei den 628 Patienten ohne Angaben zum BMI mit 19,9% doppelt so hoch wie bei den 2.837 Patienten mit erfasstem BMI.

**Ergebnisse für Modell 3 (fehlende Werte teilweise einbezogen):**

Im nächsten Schritt kann nun der einzige noch verbliebene Parameter mit fehlenden Werten, der Body Mass Index, aus dem Modell entfernt werden, da im Gesamtmodell – wie auch schon im ersten Modell (Tabelle 4.7) – keine Signifikanz hinsichtlich der Zielgröße vorlag. Somit kann im folgenden Modell (Tabelle 4.10) schließlich die volle Patientenzahl betrachtet werden.

**Tabelle 4.10: Logistische Regression 3 – soziodemographische Variablen bzgl. Krankenhaus-Letalität, mit fehlenden Werten (alle Patienten)**

Parameter	Anzahl Werte betrachtet	Anzahl (davon) fehlender Werte	p-Wert im Gesamtmodell
Alter [Jahre]	3.465	0	<0,0001
Geschlecht (männlich / weiblich)	3.465	0	0,0919
Nationalität (deutsch / nicht deutsch)	3.465	(101)	0,4621
Wohnort (Bln.-Ost/Bln.-West/Brb/sonst.)	3.465	(198)	0,0545
Familienstand (nicht allein / allein lebend)	3.465	(495)	0,0722
letzte berufliche Stellung (Arb./Ang./...)	3.465	(1.868)	0,1066
Gesamt-Modell*	3.465	(2.101)	<0,0001
Body Mass Index [ $kg/m^2$ ]**	2.837	628	,

\* 2.101 Beobachtungen (Patienten) beinhalten wenigstens einen fehlenden Wert

\*\* nur einzeln getestet, da fehlender Werte nicht einbeziehbar

Es zeigen sich im Vergleich zwischen den Modellen aus den Tabellen 4.7, 4.9 und 4.10 zwar keine Unterschiede bezüglich der Signifikanz-Einstufung, jedoch insbesondere bei Geschlecht oder Wohnort starke Unterschiede in den p-Werten zwischen den Modellen 2 und 3. Diese Unterschiede deuten auf Wechselwirkungen zwischen den Parametern hinsichtlich der Zielgröße hin, die auch bereits aufgrund der starken Unterschiede zwischen den einzelnen p-Werten und dem p-Wert im Gesamtmodell (Tabelle 4.7) vermutet werden konnten.

**Ergebnisse nach Modellselektion (Modelle 1-3):**

Wie oben ausgeführt, wurde das Alter des Patienten als einziger soziodemographischer Parameter identifiziert, der einen signifikanten Einfluss auf das Sterberisiko besitzt. Dieses Ergebnis wird im Folgenden mittels automatisierter Selektionsalgorithmen überprüft (vgl. Kapitel 3.2.2). Um die Robustheit der Ergebnisse zu betrachten, wurden dabei jeweils Vorwärts-, Rückwärts- und Schrittweise-Prozeduren

angewendet, mit  $f_\alpha = b_\alpha = 0,05$ . Für  $b_\alpha$  wird zwar häufig ein Wert  $> 0,05$  (z.B.  $b_\alpha = 0,1$ ) verwendet; aufgrund der großen Fallzahl wurde hier jedoch der strikere Wert von  $0,05$  gewählt.

Ergebnisse für Modell 1 (fehlende Werte eliminiert,  $n = 2.354$ ):

Bei jeder der drei angewandten Selektionsprozeduren verblieb jeweils nur das Alter der Patienten im Modell, jeweils mit  $p < 0,0001$ . Die Parameterschätzer der bereinigten Modelle sind daher für alle drei Methoden

$$\begin{aligned} \text{mit } \hat{\beta}_0 &= -6,7471 \text{ und} \\ \hat{\beta}_1 &= 0,0636 \end{aligned}$$

identisch ( $\beta_0$  bezeichnet das Absolutglied,  $\beta_1$  den Koeffizienten für das Alter  $X_1$ ). Somit wird für die demographischen Variablen für Modell 1 die folgende Beziehung angenommen:

$$\hat{P}(Y = 1 \mid X_1 = x_1) = (1 + \exp(6,7471 - 0,0636 x_1))^{-1} .$$

Für einen beispielsweise 70-jährigen Patienten ( $X_1 = 70$ ) errechnet sich somit ein Mortalitätsrisiko von 9,15%.

Ergebnisse für Modell 2 (fehlende Kategorien einbezogen,  $n = 2.837$ ):

Auch bei diesem vergrößerten Datensatz, bei dem fehlende Werte für nominale Einflussgrößen als eine zulässige Ausprägung berücksichtigt wurden, verblieb jeweils nur das Alter der Patienten im Modell, jeweils mit einem p-Wert von  $p < 0,0001$  (siehe Anhang). Die Parameterschätzer der reduzierten Modelle sind daher auch hier für alle drei betrachteten Methoden

$$\begin{aligned} \text{mit } \hat{\beta}_0 &= -6,6624 \text{ und} \\ \hat{\beta}_1 &= 0,0633 \end{aligned}$$

identisch.

Somit gilt für Modell 2 die gleiche Modellstruktur wie für Modell 1, nämlich mit dem Absolutglied und lediglich einem Regressionskoeffizienten, dem Alter. Für den 70-jährigen Beispielpatienten errechnet sich nach Modell 2 ein leicht erhöhtes Mortalitätsrisiko von 9,70%. Dieser höhere Wert gilt für den gesamten Altersbereich und ist in der etwas höheren Zahl der Todesfälle begründet, die in Modell 2 einbezogen werden konnte.

Ergebnisse für Modell 3 (fehlende Kategorien einbezogen, BMI eliminiert,  $n = 3.465$ ):

Bei jeder der drei angewandten Prozeduren verblieben in diesem vollständigen Datensatz, bei dem alle Todesfälle einbezogen wurden, jedoch kein BMI betrachtet wurde, neben dem Absolutglied die folgenden Parameter im Modell:

- $X_1$  : Alter [Jahre]  $(p < 0,0001)$
- $X_2$  : Geschlecht (männlich, weiblich)  $(p = 0,0446)$
- $X_3$  : Familienstand (nicht allein, allein, n.b.)  $(p = 0,0451)$

$$\begin{aligned} \text{mit } \hat{\beta}_0 &= -5,8652 ; \\ \hat{\beta}_1 &= 0,0603 ; \\ \hat{\beta}_2 &= \begin{cases} -0,2542 & \text{falls } X_2 = 1 \text{ („männlich“)} \\ 0 & \text{falls } X_2 = 2 \text{ („weiblich“)} ; \end{cases} \\ \hat{\beta}_3 &= \begin{cases} -0,4061 & \text{falls } X_3 = 0 \text{ („nicht allein lebend“)} \\ -0,2451 & \text{falls } X_3 = 1 \text{ („allein lebend“)} \\ 0 & \text{falls } X_3 = 9 \text{ („nicht erfasst“)} . \end{cases} \end{aligned}$$

Ein männlicher 70-jähriger Patient ohne erfassten Familienstand würde also ein Sterberisiko von

$$\begin{aligned} \hat{P}(Y = 1 | X_1 = 70, X_2 = 1, X_3 = 9) &= (1 + \exp(5,8652 - 0,0603 * 70 + 0,2542 + 0))^{-1} \\ &= 0,1303 \end{aligned}$$

aufweisen.

Die Signifikanz der beiden Parameter Geschlecht und Familienstand ist im Gegensatz zum Alter deutlich schwächer ausgeprägt und liegt nur knapp unterhalb der Schwelle von 5%. Zudem zeigt sie sich nur in Modell 3, dem Modell mit der höchsten Fallzahl.

Die Betrachtung des singulären Odds Ratios beim Geschlecht,

$$\widehat{\text{OR}}('m' \text{ vs. } 'w') = \frac{\frac{\hat{P}(Y=1|X_2=1)}{1-\hat{P}(Y=1|X_2=1)}}{\frac{\hat{P}(Y=1|X_2=2)}{1-\hat{P}(Y=1|X_2=2)}} = e^{\hat{\beta}_3(X_3=1)} = e^{-0,2542} = 0,7755$$

zeigen deutliche Unterschiede in der Sterbewahrscheinlichkeit. Eine 70-jährige Beispielpatientin mit Familienstand „nicht erfasst“ hätte entsprechend eine Sterbewahrscheinlichkeit von 0,1619. Wie in Abschnitt 4.4.1.4 diskutiert wird, sind die beiden Kovariaten Alter und Geschlecht miteinander in der Weise verknüpft, dass die eingeschlossenen Frauen durchschnittlich etwa 10 Jahre älter sind als die Männer und hierbei eine höhere Sterblichkeit zeigen als durch den Faktor Alter alleine erklärt wird.

Beim Familienstand besteht die Signifikanz nur im Vergleich zwischen den beiden Ausprägungen „nicht allein lebend“ und „nicht erfasst“ ( $\widehat{\text{OR}} = 0,6662$ ;  $p = 0,0128$ ), und nicht etwa zwischen „nicht allein lebend“ und „allein lebend“. Bei der Interpretation dieses Ergebnisses tritt der bereits besprochene Umstand auf, dass ja gerade als Hauptgrund für das Fehlen von Eintragungen zur Demographie und anderen Einflussgrößen das (frühe) Versterben selbst vermutet wird. Die Kausalität ist demnach also eher vom Versterben auf das Fehlen von Werten als in umgekehrter Richtung zu sehen. Es kann weiterhin vermutet werden, dass der Familienstand mit dem Alter in der Weise zusammenhängt, dass ältere Patienten häufiger allein leben (siehe auch Kap. 4.4.1.4); außerdem könnten allein lebende Patienten die höhere Prähospitalzeit aufweisen, da keine weitere Person mit ihnen im selben Haushalt lebt. Aufgrund dessen wird auch auf diesen Parameter bei der späteren Risikoadjustierung verzichtet.

Als ein erstes konsolidiertes Modell zur Demographie wird also der gesamte Datensatz ( $n = 3.465$ ) mit  $X_1$  (Alter) und  $X_2$  (Geschlecht) als unabhängige Größen betrachtet. Hierfür gelten die folgenden Schätzparameter (mit 95%-Konfidenzgrenzen nach Wald):

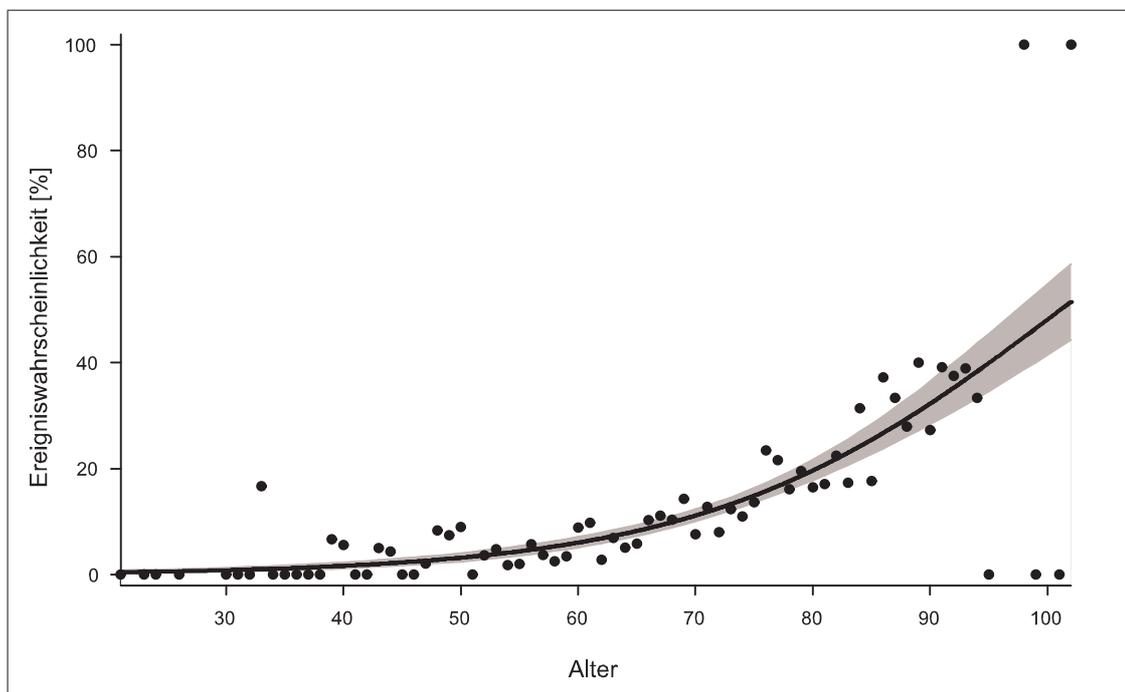
$$\begin{aligned}\hat{\beta}_0 &= -6,1968 & [-7,0046 ; -6,0253] \\ \hat{\beta}_1 &= 0,0616 & [ 0,0514 ; 0,0717] & (p < 0,0001), \text{ und} \\ \hat{\beta}_{21} &= -0,3203 & [-0,5526 ; 0,0880] & (p = 0,0069).\end{aligned}$$

Die geschätzten Odds Ratios für die beiden Parameter lauten:

$$\begin{aligned}\widehat{\text{OR}}(X_1) &= e^{(\hat{\beta}_1)} = 1,063 \\ \widehat{\text{OR}}(X_{21}) &= e^{(\hat{\beta}_{21})} = 0,726 ,\end{aligned}$$

d.h. dass sich das Letalitätsrisiko mit steigendem Lebensalter um durchschnittlich etwa 6,3% pro Lebensjahr erhöht. Männer besitzen hiernach ein um etwas mehr als ein Viertel reduziertes Chance-Risiko-Verhältnis. Der Effekt des Alters besitzt jedoch erkennbar den wichtigsten Einfluss auf das Sterberisiko. Dieser ist in Abbildung 4.5 mit der entsprechenden Schätzfunktion, dem zugehörigen Konfidenzband und den beobachteten Letalitätsraten für jedes vollendete Lebensjahr illustriert.

**Abbildung 4.5: Krankenhaus-Letalität nach Altersgruppen**



#### 4.4.1.3 CART-Analyse für demographische Variablen

Als Alternative zur statistischen/parametrischen Modellierung werden in diesem Abschnitt die Ergebnisse der im Methoden-Kapitel 2.6 vorgestellten CART-Analyse betrachtet, die weitere Aufschlüsse über Einfluss und Zusammenhang der demographischen Kovariaten gibt. Die Analyse wurde mit exakt denselben Variablen und Ausprägungen durchgeführt wie die logistische Regression.

Die Analyse, durchgeführt mit dem Programmpaket *R* (Version 2.0.1), zeigt das in Abbildung 4.6 dargestellte Ergebnis. Die benutzte Einstellung für das verwendete (Un)Reinheitsmaß (Splitregel) ist der Gini-Index (vgl. Kapitel 2.6.3.1).

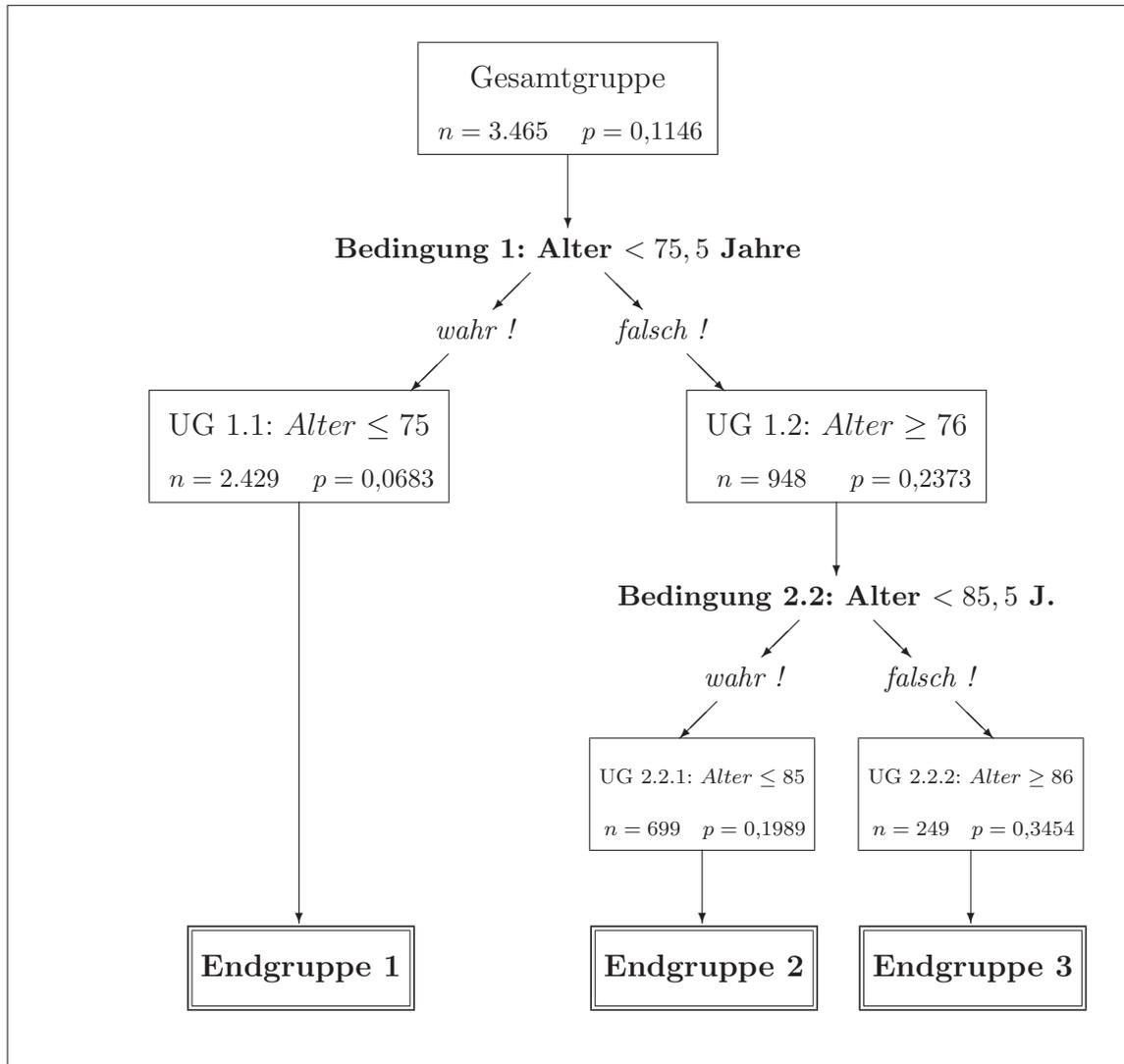
Bei der CART-Analyse zeigt sich bei den demographischen Variablen ausschließlich das Alter als relevanter Einflussfaktor hinsichtlich der Krankenhaus-Letalität. Innerhalb des Altersbereichs der betrachteten Population (20 Jahre bis 102 Jahre) wurden zwei relevante Trennstellen gefunden: die erste bei einem Alter von 75,5 Jahren, die zweite bei 85,5 Jahren. Die so gebildeten drei Endgruppen weisen hinsichtlich des Gini-Index zueinander die größte und innerhalb die kleinste Unreinheit auf.

Abbildung 4.7 zeigt die beobachteten Todesraten nach Lebensalter in Jahren und die vom CART-Algorithmus bestimmten drei Endgruppen mit den zugehörigen Letalitätsraten. Zusätzlich sind die vom logistischen Regressionsmodell (nur Alter als Einflussfaktor) geschätzten Todeswahrscheinlichkeiten mit folgenden Parameterschätzungen (vgl. Seite 171) eingezeichnet:

$$\begin{aligned}\hat{\beta}_0 &= -6,7398 \quad (\text{Intercept}), \\ \hat{\beta}_1 &= +0,0667 \quad (\text{Slope}).\end{aligned}$$

In diesem konkreten Beispiel zeigen sich also durchaus übereinstimmende Ergebnisse der konzeptionell unterschiedlichen Methoden, zumindest hinsichtlich der als signifikanten bzw. für die Partitionierung relevanten Einflussgrößen beider Verfahren. In beiden Verfahren erweist sich lediglich das Alter der Patienten als für die Krankenhaus-Letalität signifikant/relevant, obwohl bei einigen der logistischen Modelle auch das Geschlecht einen Einfluss zeigt. Weiterhin ist ein (streng) monotoner Anstieg der Sterbewahrscheinlichkeit mit zunehmendem Alter zu vermuten.

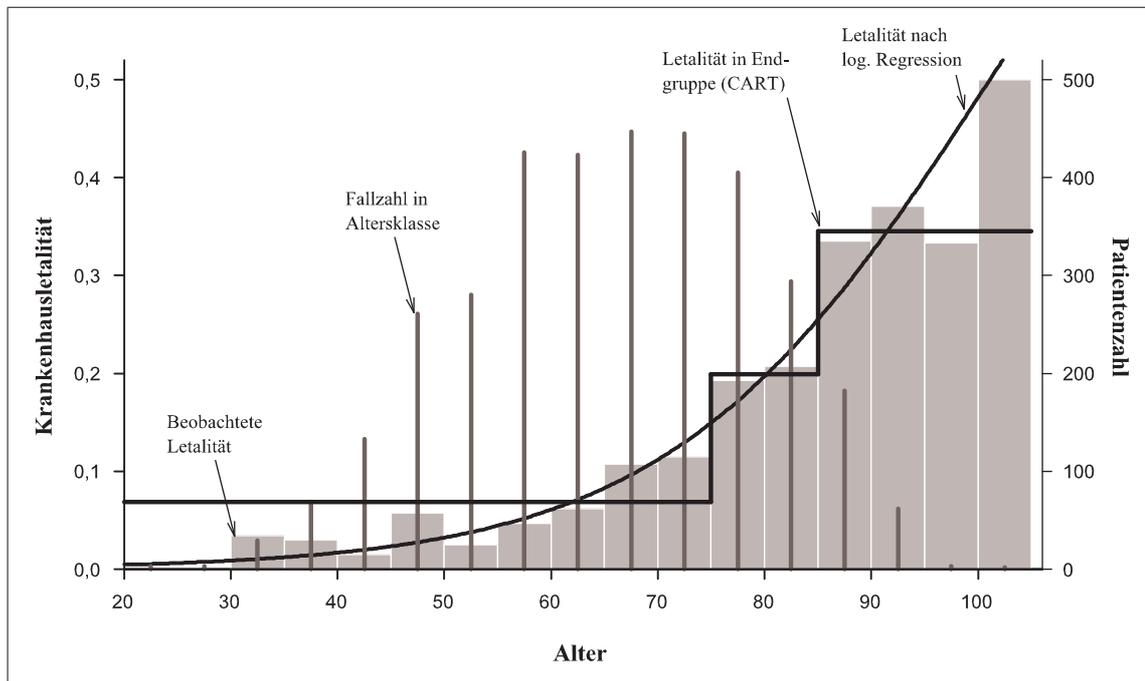
Abbildung 4.6: CART-Analyse für demographische Variablen



UG: Untergruppe  $n$ : Fallzahl in Knoten  $p$ : Letalität in Knoten

Als einschränkend für die Vergleichbarkeit der Methoden muss der Umstand angesehen werden, dass bei der CART-Methode – im Gegensatz zum linearen Modell – keine Monotonität (Linearität) des Regressors auf die Zielgröße vorausgesetzt wird. Die logistische Regression würde bei komplexeren, nicht-monotonen, funktionalen Zusammenhängen zwischen Alter und Letalität schlechtere Schätzungen liefern und eventuell vorhandene signifikante Einflüsse nicht aufdecken. Die Methodologie der CART-Verfahren ist hiervon jedoch unabhängig, da kein spezieller funktionaler Zusammenhang vorausgesetzt wird.

**Abbildung 4.7: Krankenhaus-Letalität nach Altersklassen, Daten und Analyse-Ergebnisse**



Wie in Abbildung 4.7 dargestellt ist, sind die Patientenzahlen im Altersbereich von über 95 Jahren gering ( $n = 5$ ). Gleichzeitig werden in dieser Gruppe die höchsten Letalitätsraten (40%) beobachtet. Würde man nun die Altersgruppe über 95 Jahre aus dem Datensatz eliminieren, wären die Endgruppen beim CART-Algorithmus unverändert (Endgruppe 3:  $n = 244$ ,  $p = 0,3443$ ). Beim logistischen Modell würden sich die Schätzparameter geringfügig ändern zu:

$$\begin{aligned}\hat{\beta}'_0 &= -6,75626 \quad (\text{Intercept}), \\ \hat{\beta}'_1 &= +0,06690 \quad (\text{Slope}).\end{aligned}$$

Der Schätzwert für die Letalitätswahrscheinlichkeit an der Stelle  $Alter = 88$  (Alters-Median in Endgruppe 3) von  $\hat{p} = 0,2945$  im Gesamtmodell würde nahezu unverändert bleiben ( $\hat{p}' = 0,2955$ ). Beide Verfahren zeigen sich gegenüber Daten mit kleinen Fallzahlen am (oberen) Rande des Wertebereichs des Regressors als robust.

Das Bilden von Untergruppen beim CART-Modell kann aber andererseits die Wirklichkeit wesentlich unzureichender abbilden als beispielsweise ein lineares Modell. Falls – wie im vorliegenden Beispiel beim Alter vermutet werden kann – ein

funktionaler und monotoner Zusammenhang zwischen einer (stetigen) Einflussgröße und der Zielgröße besteht, erscheint die Aussagekraft der CART-Analyse im Vergleich zum linearen Modell geringer. Dies gilt insbesondere bei nur einer verbleibenden Einflussgröße im Baum. Würden die Daten dem monotonen Zusammenhang zwischen Einfluss- und Zielgröße sehr genau folgen, so wäre die Partitionierung der Population eher instabil und würde wenig Information liefern.

Um die Eigenschaften der einzelnen bzw. die Zusammenhänge zwischen den demographischen Kovariaten weiter zu beleuchten, wird im Folgenden eine separate Betrachtung der oben modellierten Einflussfaktoren zusammen mit der Zielgröße durchgeführt.

Einzeln betrachtet zeigen die meisten demographischen Merkmale Signifikanz hinsichtlich der Letalität (nur der Wohnort nicht), jedoch gilt dies im Gesamtmodell lediglich für das Alter des Patienten. Die p-Werte für diese Merkmale liegen im Gesamtmodell deutlich im nicht-signifikanten Bereich, wobei hingegen der p-Wert für den Faktor Wohnort kaum verändert ist. Dieser Umstand deutet darauf hin, dass zumindest zwischen diesen Kovariaten und dem Alter ein (zumindest bivariater) Zusammenhang („Kollinearität“) besteht und dass die unterschiedliche Letalität ausschließlich durch Unterschiede beim Alter erklärt werden kann.

Die folgenden explorativen Betrachtungen geben über diese Vermutung Aufschluss.

#### **4.4.1.4 Univariate Betrachtungen demographischer Kovariaten**

##### **a) Alter**

Der (monotone) Zusammenhang zwischen dem Alter und der Krankenhaus-Letalität wurde bereits in obiger Darstellung gezeigt. Die Sterblichkeit nach Altersklassen wird zusätzlich in Tabelle 4.12 dargestellt.

##### **b) Geschlecht**

Die Gruppe der Frauen zeigt im Vergleich zu der der Männer eine etwas mehr als doppelt so hohe Letalität (17,7% gegen 8,1%, siehe Tabelle 4.12). Das Relative Risiko (RR) liegt bei 2,19 (95%-KI: [1,82 ; 2,63]) und das Odds Ratio (OR) bei 2,44

(95%-KI: [1,98 ; 3,02]). Jedoch weist die obige Modellbetrachtung darauf hin, dass das Geschlecht im Gesamtmodell (also zusammen mit dem Alter) einen schwächeren Einfluss besitzt, da die Ursache für diesen Unterschied teilweise in dem höheren Alter der Frauen beim Infarkt, und nicht ausschließlich im Geschlecht des Patienten, liegt.

**Tabelle 4.11: Letalität nach Altersgruppen und Geschlecht**  
– vollständige Fälle (n=2.354)

Geschlecht	Gesamt n Letal.	20 - 29 J. n Letal.	30 - 39 J. n Letal.	40 - 49 J. n Letal.	50 - 59 J. n Letal.	60 - 69 J. n Letal.	70 - 79 J. n Letal.	80 - 89 J. n Letal.	über 90 J. n Letal.
männlich	1572 7,0%	2 0,0%	43 0,0%	226 3,1%	396 2,5%	435 8,3%	326 8,9%	122 19,7%	22 18,2%
weiblich	782 13,3%	1 0,0%	11 9,1%	29 3,4%	99 2,0%	152 7,9%	243 14,4%	217 19,4%	30 36,7%
Gesamt	2354 9,1%	3 0,0%	54 1,9%	255 3,1%	495 2,4%	587 8,2%	569 11,2%	339 19,5%	52 28,8%

Wie vermutet, zeigen sich nur geringe Letalitäts-Unterschiede zwischen den Geschlechtern innerhalb der Altersgruppen, sehr wohl aber zwischen den Altersgruppen. Betrachtet man jedoch alle verfügbaren Fälle für Alter und Geschlecht, zeigt sich ein unterschiedliches Bild (vgl. Tabelle 4.12). In dieser Darstellung ist die weibliche Letalität in allen Altersgruppen größer als die männliche.

**Tabelle 4.12: Letalität nach Altersgruppen und Geschlecht**  
– alle verfügbaren Fälle (n=3.465)

Geschlecht	Gesamt n Letal.	20 - 29 J. n Letal.	30 - 39 J. n Letal.	40 - 49 J. n Letal.	50 - 59 J. n Letal.	60 - 69 J. n Letal.	70 - 79 J. n Letal.	80 - 89 J. n Letal.	über 90 J. n Letal.
männlich	2251 8,1%	4 0,0%	65 1,5%	297 3,4%	566 3,7%	635 8,2%	479 11,3%	176 20,5%	29 27,6%
weiblich	1214 17,7%	1 0,0%	14 7,1%	48 4,2%	128 3,9%	222 9,9%	384 18,0%	357 26,1%	60 38,3%
Gesamt	3465 11,5%	5 0,0%	79 2,5%	345 3,5%	694 3,7%	857 8,6%	863 14,3%	533 24,2%	89 34,8%

Das logistische Modell, welches als Einflussfaktoren das Alter und das Geschlecht betrachtet und alle 3.465 Fälle einschließt, zeigte Signifikanz auch für den Faktor Geschlecht ( $p = 0,0069$ ). Das gleiche Modell mit den hinsichtlich der soziodemographischen Faktoren 2.265 vollständigen Fällen zeigte für den Faktor Geschlecht einen p-Wert von 0,2799, also wie im Gesamtmodell keine Signifikanz. Wegen der starken Signifikanz im erstgenannten Modell sollte das Geschlecht als Kovariate für die spätere Adjustierung verwendet werden.

Dieses Beispiel verdeutlicht die möglichen Schwierigkeiten bei der Interpretation von komplexen Datensätzen empirischer Erhebungen, die eine hohe Zahl von

fehlenden Werten beinhalten. Es liegen häufig starke Abhängigkeiten zwischen den Einflussfaktoren und zusätzlich mögliche systematische (also nicht zufällige) Ausfälle bzw. fehlende Werte vor, wie hier durch die deutlich geringere Letalität der vollständigen Fälle und der Gesamt-Stichprobe gezeigt wird (vgl. Tabelle 4.8).

### c) Body Mass Index

Ähnlich wie beim Geschlecht zeigt der BMI im univariaten Modell eine Signifikanz hinsichtlich des Sterberisikos. Tabelle 4.13 veranschaulicht, dass die Letalität mit zunehmendem BMI monoton abnimmt. Der  $\chi^2$ -Test auf Unterschiede zwischen den BMI-Klassen (der aufgrund der Klassifizierung nicht dem Test im Gesamtmodell entspricht) – wie im Modell unter Berücksichtigung der Rohwerte (siehe Tabelle 4.7) – liefert einen signifikanten p-Wert ( $p = 0,0044$ ), so dass auf Letalitätsunterschiede zwischen diesen Klassen geschlossen werden kann.

**Tabelle 4.13: Letalität nach BMI-Klassen**  
– alle verfügbaren Fälle

BMI-Klasse	n	Letalität
1 „untergewichtig“: BMI < 20	104	14,42%
2 „normal“: $20 \leq \text{BMI} < 25$	990	11,62%
3 „übergewichtig“: $25 \leq \text{BMI} < 30$	1.228	8,71%
4 „adipös“: $30 \leq \text{BMI}$	515	6,80%
Gesamt	2.837	9,59%
BMI unbekannt	628	19,90%

Wie bereits vorher diskutiert, zeigt sich auch hier in der Gruppe der Patienten ohne verfügbaren BMI die weitaus höchste Letalität (125 Fälle oder 19,9%). Die Herzinfarktsterblichkeit, wenn überhaupt das Klinikum erreicht worden ist, ist – mit den Einschränkungen bei fehlenden Werten – offensichtlich mit höherem BMI rückläufig; die Gruppe der Untergewichtigen zeigt etwa im Vergleich zu den adipösen Patienten ein Odds Ratio von 2,31 (95%-KI: [1,21 ; 4,41]) und ein relatives Risiko von 2,12 (95%-KI: [1,20 ; 3,74]). Entsprechend der allgemein akzeptierten Annahme steigt jedoch das Herzinfaktrisiko und auch die Gesamtsterblichkeit mit zunehmendem BMI an.

Wie für das Geschlecht, verschwindet auch für den BMI die Signifikanz hinsichtlich der Sterblichkeit bei der Hinzunahme des Alters ins Modell. Bei allen 2.837 Fällen, für die Alter und BMI erhoben sind (zwei Fälle wurden wegen fraglicher Plausibilität [BMI>60] ausgeschlossen), ergibt sich im Modell mit beiden Faktoren ein p-Wert von 0,7445 für den BMI (bzw.  $p = 0,7078$  für alle 2.354 vollständigen Fälle).

Die Abnahme des BMI in den oberen Altersklassen kann im Zusammenhang mit der Zunahme der Letalität in diesen Altersklassen gesehen werden. Die Annahme, dass der BMI in der betrachteten Population einen eigenen Beitrag zur Erklärung der Letalitäts-Wahrscheinlichkeit liefert, darf somit bezweifelt werden. Ein positiver Zusammenhang zwischen BMI und der Wahrscheinlichkeit für einen Infarkt in der Zukunft (auch nach einer eventuellen Gewichtsreduktion) sowie der Letalitäts-wahrscheinlichkeit nach dem Infarkt bzw. nach dem Erreichen der Klinik gilt als medizinisch gesichert. Es stellt sich daher die Frage, ob der BMI in die später zu erfolgende Risikoadjustierung einbezogen werden sollte, obwohl dieses Merkmal keinen erkennbaren Beitrag zur Letalität liefert. Fragen dieser Art können nur in Zusammenarbeit mit Experten endgültig beantwortet werden (vgl. Kapitel 4.4.2).

#### **d) Nationalität**

Die Gruppe der Deutschen zeigt im Vergleich zu den Ausländern eine knapp doppelt so hohe Letalität (11,7% gegen 6,3%). Das Relative Risiko (RR) liegt bei 1,86 (95%-KI: [1,07 ; 3,25]) und das Odds Ratio (OR) bei 1,98 (95%-KI: [1,09 ; 3,59]). Jedoch weist die Modellbetrachtung aus Tabelle 4.7 darauf hin, dass die Nationalität im Gesamtmodell (also zusammen mit dem Alter) – ähnlich wie beim BMI – keinen Einfluss besitzt, sondern dass die Ursache für diesen Unterschied in dem höheren Alter beim Infarkt liegt. Tatsächlich ist die Sterblichkeit innerhalb der Altersgruppen zwischen den Nationalitätsgruppen in etwa gleich.

Zur Signifikanzprüfung wurde wieder ein logistisches Modell angepasst, welches als Einflussfaktoren lediglich das Alter und die Nationalität beinhaltet, um so alle für Nationalität vollständigen 3.364 Fälle einschließen zu können. Der p-Wert für Nationalität liegt in diesem Modell bei 0,7587, zeigt also wie im Gesamtmodell keine Signifikanz. Zum Vergleich: Das gleiche Modell mit den (hinsichtlich der soziodemographischen Faktoren) 2.265 vollständigen Fällen zeigte für den Faktor Nationalität einen p-Wert von  $p = 0,2722$ , also wie im Gesamtmodell keine Signifikanz.

**Tabelle 4.14: Letalität nach Nationalität**  
– alle verfügbaren Fälle

Nationalität	n	Letalität
deutsch	3.172	11,66%
nicht deutsch	192	6,25%
Gesamt	3.364	11,36%
Nationalität unbekannt	101	14,85%

Die geringere Letalität der nicht deutschen Patienten ist also in Zusammenhang mit dem geringeren Alter dieser Gruppe zu sehen, was die Beibehaltung der Nationalität im Modell als nicht erforderlich erscheinen lässt.

#### d) Wohnort

Beim Wohnortes des Patienten kann vermutet werden, dass dieser mit der – für die Prognose des Patienten wichtigen – Prähospitalzeit zusammenhängt. Patienten, die im Umland und damit weiter von der nächstgelegenen Notfallklinik entfernt wohnen, könnte die Prähospitalzeit höher sein, als bei Patienten aus dem Berliner Stadtgebiet.

**Tabelle 4.15: Letalität nach Wohnort**  
– alle verfügbaren Fälle

Wohnort	n	Letalität
Berlin-Ost	666	9,91%
Berlin-West	2.381	11,97%
Brandenburg	108	4,63%
sonstige	112	11,61%
Gesamt	3.267	11,36%
Wohnort unbekannt	198	14,14%

Bei Betrachtung der Mortalität in den vier betrachteten Gruppen kann diese Vermutung jedoch nicht belegt werden. Im Gegenteil ist die Sterblichkeit der Brandenburger Patienten weitaus geringer als die der übrigen Regionen. Wie im folgenden

Abschnitt gezeigt wird, kann dies wiederum auf das geringere Alter der Patienten aus Brandenburg zurückgeführt werden.

Die Prähospitalzeit ist in der vorliegenden Datenbank sehr lückenhaft erfasst (Anteil fehlender Werte: 28,1%), so dass der Zusammenhang zum Wohnort und zur Mortalität nicht valide zu bewerten ist.

Ähnliche Zusammenhänge liegen zwischen dem Alter und den übrigen demographischen Variablen vor, die daher hier nicht weiter im Detail beschrieben sind.

#### **4.4.1.5 Bivariate Betrachtungen demographischer Kovariaten mit dem Patientenalter**

Dem Alter der Patienten kommt laut erster Modellbetrachtungen (logistische Regressionsanalyse und CART-Analyse) bei den demographischen Kovariaten offensichtlich die entscheidende Rolle hinsichtlich der Letalität zu. Die Signifikanzen, die die übrigen Parameter in den Einzelbetrachtungen – jedoch weniger im Gesamtmodell – zeigen, sind somit zu einem großen Teil auf die Kollinearität, d.h. die Abhängigkeit zwischen diesen Parametern und dem Alter, zurückführbar.

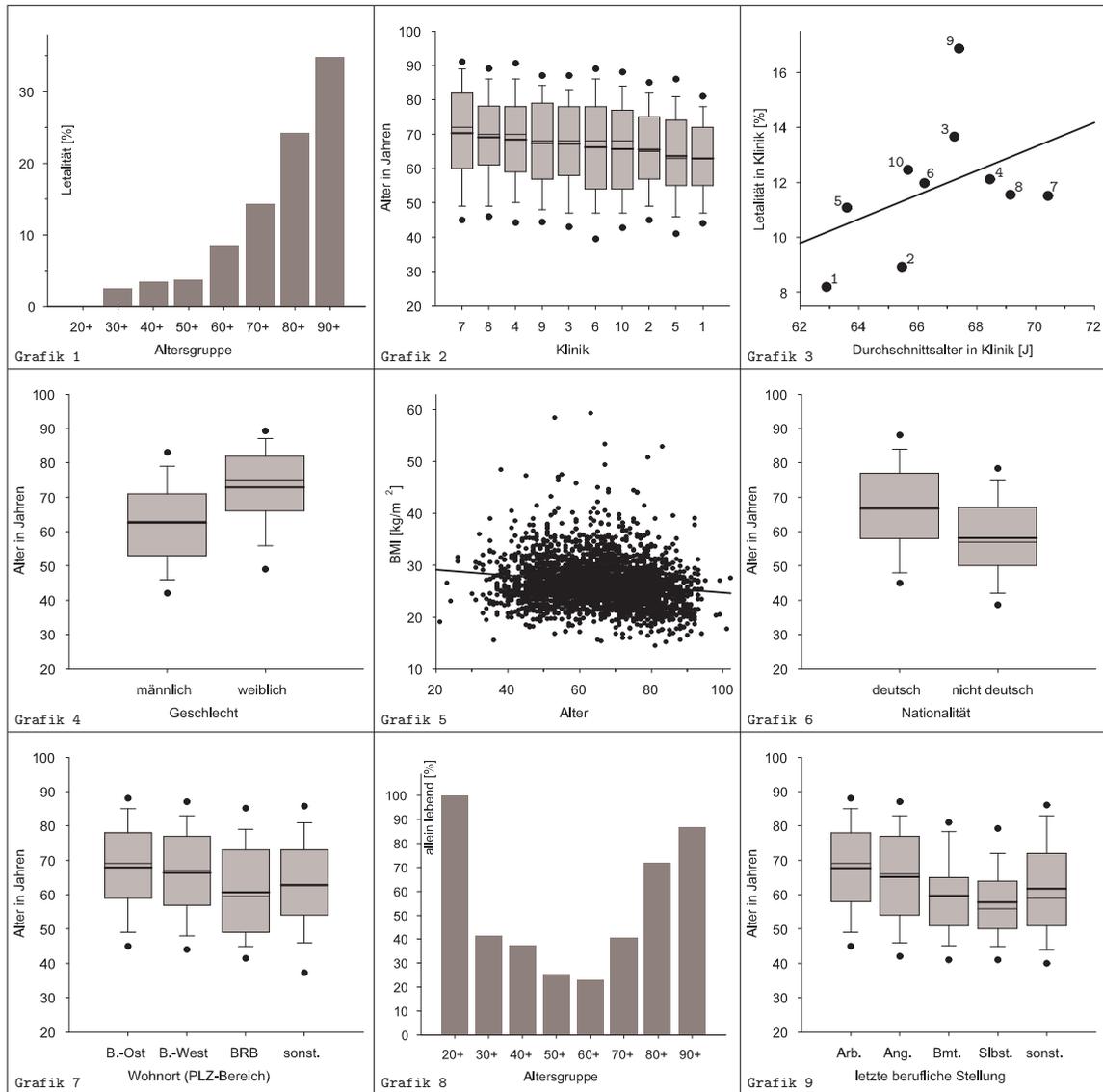
In Abbildung 4.8 sind bivariate Betrachtungen der Schlüsselvariablen „Alter“ mit der Zielgröße, der Klinik und weiteren soziodemographischen Variablen dargestellt.

##### **1. Letalität und Alter**

Die erste Grafik in Abbildung 4.8 veranschaulicht den Zusammenhang des Alters und der Letalitätsraten aus obigen Modellen. Die Sterblichkeit der Patienten beispielsweise über 80 Jahre ist mit 26,7% etwa 8-mal so hoch wie in der Gruppe der unter 50-jährigen Patienten (3,3%).

Weiterhin zeigen sich Unterschiede zwischen den Zentren hinsichtlich des Alters (zweite Grafik). Zentrum Nr. 7 ist dasjenige mit dem höchsten Durchschnittsalter (70,4 Jahre); Klinikum Nr. 1 schloss mit einem Durchschnittsalter von 62,9 Jahren die jüngsten Patienten in die Erhebung ein und zeigt gleichzeitig die geringste Letalität (vgl. Tabelle 4.3). Die Auswertung einer univariaten ANOVA mit „Klinikum“ als Einflussgröße zeigte hinsichtlich des Alters Unterschiede zwischen den Kliniken ( $p < 0,0001$ ).

Abbildung 4.8: Zusammenhang des Alters mit demographischen Kovariaten



Die dritte Grafik in Abbildung 4.8 zeigt die zehn Datenpunkte mit Durchschnittsalter in der Klinik und der beobachteten Letalität. Die Steigung der eingezeichneten Regressionsgeraden verdeutlicht, dass Kliniken mit älterem Patientenkollektiv auch eher höhere Letalitätsraten aufweisen.

## 2. Geschlecht und Alter

Die vierte Grafik in Abbildung 4.8 zeigt, dass die Gruppe der Männer – bei gleicher Streuung – im Mittel etwa zehn Jahre jünger ist als die Gruppe der Frauen (siehe auch Tabelle 4.16). Beim Vergleich zwischen den beiden Geschlechtsgruppen zeigen sich mittels t-Tests – und ebenfalls mittels Wilcoxon-Tests – hochsignifikante Unterschiede hinsichtlich des Alters ( $p < 0,0001$ ).

**Tabelle 4.16: Alter (in Jahren) nach Geschlecht**

Geschlecht	Anzahl	Mittelwert	St.Abweich.	Minimum	Median	Maximum
männlich	2.251	62,54	12,52	21	63	101
weiblich	1.214	72,95	12,45	26	75	102
Gesamt	3.465	66,18	13,45	21	67	102

## 3. Body Mass Index und Alter

Der Scatter-Plot in Abbildung 4.8 (fünfte Grafik) zeigt die BMI-Verteilung in Abhängigkeit des Alters. Die eingezeichnete Regressionsgerade deutet auf eine leichte Abnahme des BMI bei höherem Alter hin.

Dieser Zusammenhang wird durch die beobachteten bivariaten Korrelationen belegt. Der Korrelationskoeffizient nach Pearson berechnet sich für  $n = 2.354$  zu  $r = -0,1649$ , und für alle  $n = 2.837$  verfügbaren Fälle zu  $r = -0,1648$ . Die Nullhypothese auf Unabhängigkeit der beiden Merkmale wird in beiden Fällen mit  $p < 0,0001$  abgelehnt.

**Tabelle 4.17: Body Mass Index nach Altersklassen**

	Gesamt	20 - 29 J.	30 - 39 J.	40 - 49 J.	50 - 59 J.	60 - 69 J.	70 - 79 J.	80 - 89 J.	über 90 J.
n*	2.837	5	66	303	595	703	701	401	63
fehlende Werte	628	0	13	42	99	154	162	132	26
Mittelwert	26,62	26,21	27,12	26,84	27,38	27,44	26,33	24,78	24,00
Standardabw.	4,50	5,24	4,98	4,43	4,59	4,51	4,27	4,00	4,36
Median	26,17	26,59	26,32	26,23	26,81	26,83	26,06	24,61	24,22

\*n bezeichnet die Anzahl verfügbaren Werte in der jeweiligen Gruppe

Tabelle 4.17 verdeutlicht jedoch, dass sich die BMI-Abnahme im Wesentlichen auf die älteren Patientengruppen im Alter über 70 Jahre beschränkt und dass bis zur Gruppe von 60 bis 70 Jahren eher eine leichte BMI-Zunahme zu beobachten ist, was bei Ansicht des Scatter-Plots zunächst nicht sofort auffällt. Die Linearitätsannahme des parametrischen Korrelationskoeffizienten nach Pearson und der eingezeichneten einzelnen Regressionsgeraden in Grafik 7 ist somit fraglich. Daher sei zum Vergleich der Spearman'sche Rangkorrelationskoeffizient genannt, der zwar Monotonie voraussetzt, nicht jedoch die Linearität. Für die  $n = 2.354$  vollständigen Fälle ergibt sich ein Wert von  $r = -0,1796$  und für die  $n = 2.837$  verfügbare Fälle  $r = -0,1788$  und somit sehr ähnliche Werte wie der Pearson'sche Korrelationskoeffizient.

#### 4. Nationalität und Alter

Der Box-Plot in Abbildung 4.8 (sechste Grafik) sowie Tabelle 4.18 zeigt die Altersverteilung in beiden erfassten Nationalitätsgruppen „deutsch“ und „nicht deutsch“ in der Gesamtstichprobe.

**Tabelle 4.18: Alter (in Jahren) nach Nationalität**

Nationalität	Anzahl	Mittelwert	St.Abweich.	Minimum	Median	Maximum
deutsch	3.172	66,74	13,33	21	67	102
nicht deutsch	192	58,14	12,02	33	57	84
unbekannt	101	63,94	14,50	31	66	91
Gesamt	3.465	66,18	13,45	21	67	102

Die Gruppe der ausländischen Patienten ist – bei etwas geringerer Streuung – im Durchschnitt etwa acht Jahre jünger als die Gruppe der deutschen Patienten. Die Gruppe der Patienten, bei denen die Nationalität nicht erfasst wurde, liegt im mittleren Alter etwa in der Mitte zwischen diesen beiden Gruppen. Eine zusätzliche Verzerrung, bedingt durch fehlende Werte in dieser Variablen (deren Anzahl auch vergleichsweise klein ist), muss daher nicht vermutet werden. Selbst in dem Falle, dass tatsächlich alle 101 fehlenden Angaben zu ausländischen Patienten gehören würden, so wäre die Gruppe der Ausländer noch im Mittel um sechs Jahre jünger als die Gruppe der deutschen Patienten.

Beim Vergleich zwischen den beiden Gruppen zeigen sich mittels t-Test (ebenfalls mittels Wilcoxon-Test) Unterschiede hinsichtlich des Alters ( $p < 0,0001$ ).

## 5. Wohnort und Alter

Der Box-Plot in Abbildung 4.8 (Grafik 7) sowie Tabelle 4.19 zeigt die Altersverteilung in den vier betrachteten Wohnort-Gruppen in der Gesamtstichprobe.

**Tabelle 4.19: Alter (in Jahren) nach Wohnort**

Wohnort	Anzahl	Mittelwert	St.Abweich.	Minimum	Median	Maximum
Berlin-Ost	666	67,21	13,72	24	68	102
Berlin-West	2.381	66,33	13,24	21	67	99
Brandenburg	108	60,74	13,36	38	60	91
sonstige	112	62,74	14,01	23	63	93
unbekannt	198	65,89	13,75	33	66	95
Gesamt	3.465	66,18	13,45	21	67	102

Die Gruppe der Patienten, die nicht im Berliner Stadtgebiet wohnen, zeigen ein im Mittel um etwa fünf Jahre geringeres Alter als die Berliner Patienten, wobei die Patienten aus dem Berliner Umland (Bundesland Brandenburg) die im Durchschnitt jüngste Gruppe darstellen.

Beim Vergleich zwischen den vier Gruppen zeigen sich mittels F-Tests bzw. Kruskal-Wallis-Tests Unterschiede hinsichtlich des Alters ( $p < 0,0001$ ).

## 6. Familienstand und Alter

Tabelle 4.20 zeigt die Altersverteilung in beiden Gruppen „nicht allein lebend (verheiratet)“ und „allein lebend“ in der Gesamtstichprobe.

Die allein lebenden Patienten sind im Mittel etwa sieben Jahre älter als die verheirateten Patienten. Beim Vergleich zwischen beiden Gruppen ist jedoch zu beachten, dass die Verteilungen in diesen beiden Gruppen von unterschiedlichem Typ sind (siehe auch Balkendiagramm in Abbildung 4.8, Grafik 8): Der Anteil der allein Lebenden ist bei den sehr jungen und bei den sehr alten Patienten deutlich höher als

**Tabelle 4.20: Alter (in Jahren) nach Familienstand**

Familienstand	Anzahl	Mittelwert	St.Abweich.	Minimum	Median	Maximum
verheiratet	1.819	63,30	11,60	30	64	99
allein lebend	1.151	70,40	15,03	21	74	102
unbekannt	495	66,99	13,20	23	68	94
Gesamt	3.465	66,18	13,45	21	67	102

bei den Patienten in den mittleren Altersgruppen. Dieser Unterschied zeigt sich auch durch die deutlich größere Streuung in der Gruppe der allein lebenden, verglichen mit den verheirateten, Patienten.

Ein Chi-Quadrat Test auf Unterschiede zwischen den Altersgruppen in der  $2 \times 8$ -Kontingenztafel (Daten wie in Grafik 8 dargestellt) zeigt hochsignifikante Unterschiede in der Verteilung des Familienstands zwischen den Altersgruppen ( $p < 0,0001$ ).

## 7. Letzte berufliche Stellung und Alter

Tabelle 4.21 zeigt die Altersverteilung in den beiden Berufsgruppen „nicht allein lebend (verheiratet)“ und „allein lebend“ in der Gesamtstichprobe. In diese Betrachtung konnten nur Patienten einbezogen werden, bei denen im Falle einer bereits bestehenden Berentung der letzte Beruf *vor* der Berentung abgefragt wurde. Diese Abfrage wurde erst nach der Änderung des Fragebogens eingeführt. Für bereits berentete Patienten, die vor der Änderung der Erhebung erfasst wurden, liegt eine entsprechende Information folglich nicht vor.

Die verschiedenen Berufsgruppen zeigen kaum Unterschiede hinsichtlich ihres Alters, lediglich die Gruppe der Selbständigen ist im Mittel etwa drei Jahre jünger als die Patienten in den übrigen Gruppen. Die Gruppe, bei denen die letzte berufliche Stellung unbekannt ist, zeigt mit etwa 72 Jahren jedoch einen deutlich höheren Altersdurchschnitt. Hierfür dürfte hauptsächlich der Umstand ursächlich sein, dass bei berenteten Patienten, die naturgemäß ein höheres Alter aufweisen als die (noch) nicht berenteten Patienten, vor der Änderung des Fragebogens nach der letzten beruflichen Stellung gefragt wurde.

**Tabelle 4.21: Alter (in Jahren) nach letzter beruflicher Stellung**

Berufsgruppe*	Anzahl	Mittelwert	St.Abweich.	Minimum	Median	Maximum
Arbeiter	319	60,33	12,65	23	59	90
Angestellte	713	59,28	13,21	21	59	101
Beamte	100	59,65	11,76	26	60	98
Selbständige	187	57,78	11,11	31	56	92
sonstige	278	61,69	14,22	32	59	102
unbekannt	1.868	71,68	11,19	24	72	99
Gesamt	3.465	66,18	13,45	21	67	102

\*falls berentet, zuletzt ausgeübter Beruf

Beim Vergleich zwischen den Berufsgruppen (ohne fehlende Angaben) zeigen sich mittels F-Tests (Kruskal-Wallis Tests) mit  $p = 0,0164$  ( $p = 0,0465$ ) zwar Signifikanzen, jedoch fallen diese deutlich schwächer aus als bei den übrigen gestesteten demographischen Gruppen und Variablen. Aufgrund der besprochenen Schwierigkeiten bei der Erhebung dieses Parameters wird auf die weitere Betrachtung dieses Parameters verzichtet.

#### 4.4.1.6 Schlussfolgerung

Die Betrachtungen der demographischen Parameter zeigt, dass für eine Modellerweiterung das Alter und das Geschlecht der Patienten in Frage kommen. Alle übrigen Parameter zeigen zwar gewisse Einflüsse auf die Zielgröße, zumeist jedoch aufgrund von bivariaten Wechselwirkungen mit dem Alter der Patienten. Als weiteres Argument gegen die Einbeziehung anderer Parameter muss hier die teils sehr hohe Rate fehlender Angaben angeführt werden.

Auf die Betrachtung weiterer multivariater Wechselwirkungen kann im Rahmen dieser Arbeit verzichtet werden, da die in diesem Abschnitt dargestellten Datenexplorationen aufgrund der später verfügbar gewordenen fachlichen Information (Abschnitt 4.4.2) für die Hauptanalyse nicht verwendet werden.

### 4.4.2 Auswahl nach fachlichen Gesichtspunkten

Wie eingangs des Abschnitts 4.4 beschrieben, wird die Variablenauswahl zur Risikoadjustierung für die Hauptanalyse auf Basis der klinischen Gesichtspunkte durchgeführt. Für die zitierte Analyse der BHiR-Daten aus den Jahren 2004 und 2005 wurden die folgenden Variablen ausgewählt:

**Demographie:** Alter, Geschlecht (weiblich), BMI ( $< 20$  oder  $> 30$ );

**Präexistierende Risikofaktoren und Erkrankungen:** Raucher, Hypercholesterinämie (HCE), Arterielle Hypertonie (AHT), Diabetes Mellitus (DM);

**Erschwerende Nebendiagnosen:** Zustand nach früherem Infarkt (Z.n.I), Manifeste Herzinsuffizienz (MHI), Niereninsuffizienz (NI), Zustand nach PTCA (Perkutane Transluminale Koronar-Angiographie), Zustand nach Bypass-OP (ACB-OP);

**Akutdiagnostik:** Ejektionsfraktion (EF,  $< 55\%$  oder  $< 35\%$ ), ST-Hebungs-Infarkt (STEMI), Linksschenkelblock (LSB);

**Akuttherapie:** Begleittherapie (Reanimation, Defibrillation oder Einsatz einer IABP [intraaortale Ballonpumpe]).

Wechselwirkungen zwischen diesen Parametern hinsichtlich der Sterblichkeit wurden in dieser Auswahl nicht benannt.

Ein naheliegender Ansatz wäre es, für den dieser Arbeit zugrunde liegenden Datensatz dieselben Kovariaten zu verwenden wie für die beschriebene. Allerdings sind hierbei einige Einschränkungen zu beachten.

- Die Häufigkeit fehlender Eingaben ist in der vorliegenden Datenbasis bei einigen der genannten Variablen teilweise sehr hoch. Da bei den verwendeten Modellen vollständige Daten vorausgesetzt werden, kann die Zahl der nicht in die Analyse einschließbaren Patienten dadurch sehr groß werden und die Aussagekraft einschränken.

- Die Angaben zur Akutdiagnostik sind innerhalb der ersten 48 Stunden nach Einlieferung zu erfassen. Sie sind somit für bereits zuvor verstorbene Patienten nicht mehr zu ermitteln und daher bei weitem nicht vollständig erfasst.
- Die Erfassung einiger Variablen (wie Zustand nach PTCA, Zustand nach Bypass-Op., ST-Hebungsinfarkt, Einsatz einer Ballonpumpe) wurde erst nach dem Wechsel des Fragebogens (vgl. Abschnitt 4.2.1) eingeführt und kann somit an dieser Stelle nicht verwendet werden. Die Begleittherapie – im Folgenden als „Reanimationspflicht“ bezeichnet – umfasst somit im Folgenden die Kriterien Reanimation und Defibrillation.

Zur weiteren Beschreibung des Infarkt-Schweregrads wurden Angaben über pulmonale Stauungen [ja/nein] und über das Vorliegen eines kardiogenen Schocks [ja/nein] – d.h. nicht messbarer Blutdruck ODER (ein systolischer Blutdruck von nicht mehr als 100 mmHG UND eine Herzfrequenz von mindestens 100) – in der Datenbank abgelegt.

In den Tabellen 4.22 und 4.23 ist die Verteilung der genannten und verfügbaren Risikofaktoren aus der obigen Darstellung über die Kliniken, deren Einfluss auf die Mortalität und die Zahl der fehlenden Werte dargestellt.

Die p-Werte<sup>1</sup> und Odds Ratios wurden jeweils in einem separaten logistischen Regressionsmodell berechnet. In einem Modell mit allen Parametern wären nur weniger als 10% der Patientendaten vollständig und somit verwendbar. Die %-Angaben beziehen sich jeweils auf alle Fälle mit Angaben. Es zeigt sich Signifikanz hinsichtlich der Sterblichkeit in nahezu allen betrachteten Risikofaktoren. Einige Faktoren weisen ein Odds Ratio von deutlich weniger als 1 auf, was eine verringerte Sterblichkeit bei Vorhandensein des Merkmals anzeigt. Hierbei unberücksichtigt bleiben die Abhängigkeiten zwischen den Risikofaktoren, die etwa bei der zitierten Datenanalyse dadurch erkennbar wurden, dass die Signifikanz im Gesamtmodell nur noch für die Parameter Alter, Hypertonie, Hypercholesterinämie, Niereninsuffizienz, manifeste Herzinsuffizienz, Ejektionsfraktion <35% und Reanimationspflicht Bestand hatten. Alle übrigen Parameter verfehlten die Signifikanz-Schwelle von 5% teilweise deutlich und zeigten meist schwächere Odds Ratios.

Zudem fällt auf, dass Unterschiede zwischen den Kliniken bei schlecht erfassten Variablen häufig besonders stark ausgeprägt sind. Dieser Umstand lässt darauf

**Tabelle 4.22: Risikofaktoren (1) nach Klinikum**

		Alter	Frauen	BMI<20	BMI>30	Rauch.	HCE	AHT	DM
	p-Wert <sup>1</sup>	<0,001	<0,001	0,091	0,018	<0,001	<0,001	0,025	<0,001
	OR	1,07	2,45	1,62	0,64	0,38	0,44	1,30	1,96
Klinik	p-Wert <sup>2</sup>	<0,001	<0,001	0,044	0,520	<0,001	<0,001	<0,001	<0,001
gesamt	MW/%	66,18	35,04%	3,67%	18,15%	38,65%	40,81%	62,79%	26,25%
(n=3465)	fehlend	0,0%	0,0%	18,1%		1,5%			
1	MW/%	62,89	24,27%	3,37%	18,40%	42,14%	55,79%	61,13%	27,30%
(n=342)	fehlend	0,0%	0,0%	4,7%		1,5%			
2	MW/%	65,46	30,83%	1,70%	18,26%	40,44%	48,03%	59,40%	24,64%
(n=639)	fehlend	0,0%	0,0%	26,3%		0,9%			
3	MW/%	67,24	35,90%	4,29%	21,19%	39,38%	33,19%	66,59%	27,21%
(n=454)	fehlend	0,0%	0,0%	7,5%		0,4%			
4	MW/%	68,45	39,01%	7,37%	13,16%	38,12%	47,09%	60,54%	22,87%
(n=223)	fehlend	0,0%	0,0%	14,8%		0,0%			
5	MW/%	63,59	30,89%	3,59%	18,14%	44,57%	34,44%	53,59%	17,50%
(n=560)	fehlend	0,0%	0,0%	15,4%		3,0%			
6	MW/%	66,22	38,76%	4,70%	16,11%	37,88%	46,46%	64,14%	26,26%
(n=209)	fehlend	0,0%	0,0%	28,7%		5,3%			
7	MW/%	70,42	44,00%	4,62%	20,51%	31,31%	40,91%	76,26%	36,36%
(n=200)	fehlend	0,0%	0,0%	2,5%		1,0%			
8	MW/%	69,14	43,49%	4,97%	15,47%	29,64%	29,04%	70,66%	32,04%
(n=338)	fehlend	0,0%	0,0%	46,4%		1,2%			
9	MW/%	67,40	37,08%	1,74%	16,96%	32,32%	39,16%	66,92%	27,38%
(n=267)	fehlend	0,0%	0,0%	13,9%		1,5%			
10	MW/%	65,67	41,20%	3,48%	18,91%	40,95%	37,07%	62,07%	32,76%
(n=233)	fehlend	0,0%	0,0%	13,8%		0,4%			

<sup>1</sup> p-Wert auf Signifikanz (einzeln geprüft)

<sup>2</sup> p-Wert auf Unterschiede zwischen den Kliniken (Kruskal-Wallis-Test bzw.  $\chi^2$ -Test)

schließen, dass die beobachteten Raten der Risikofaktoren auch von der Dokumentationsqualität der Kliniken abhängen. Variablen mit einem Anteil fehlender Angaben von über 10% werden in der Analyse daher nicht berücksichtigt.

Für die Bildung des finalen Modells sollen die in Frage kommenden Variablen nun schrittweise eingeschlossen werden (vgl. Kapitel 3.4). Dabei wird die Verbesserung der Modellgüte bei Hinzunahme eines Parameters bzw. einer Parametergruppe betrachtet. Diese kann einerseits durch die Kapitel 2.2.4 (Seite 53) dargestellten Gütekriterien untersucht werden. Für die Zielstellung dieser Arbeit – dem Einrich-

Tabelle 4.23: Risikofaktoren (2) nach Klinikum

		Z.n.I.	MHI	NI	EF<55	EF<35	LSB	Reani.	Stau	Schock
	p-Wert <sup>1</sup>	<0,001	<0,001	<0,001	<0,001	0,025	<0,001	<0,001	<0,001	<0,001
	OR	1,88	5,01	3,47	6,93	5,40	2,85	8,70	4,51	6,64
Klinik	p-Wert <sup>2</sup>	<0,001	<0,001	<0,001	<0,001	0,024	0,001	0,152	<0,001	0,002
gesamt	Anteil	17,59%	6,36%	5,22%	64,12%	21,25%	5,89%	3,86%	26,17%	4,63%
(n=3465)	fehlend	6,0%	6,1%	6,0%	69,0%		6,4%	2,0%	0,6%	5,2%
1	Anteil	22,89%	5,12%	7,83%	63,96%	18,51%	4,45%	3,81%	14,03%	3,54%
(n=342)	fehlend	2,9%	2,9%	2,9%	9,9%		1,5%	0,3%	2,0%	0,9%
2	Anteil	13,13%	4,66%	6,72%	73,59%	22,08%	4,54%	3,17%	26,34%	4,40%
(n=639)	fehlend	9,4%	9,2%	9,2%	63,8%		6,9%	6,1%	0,8%	14,7%
3	Anteil	19,73%	6,50%	4,71%	48,39%	9,68%	6,25%	3,30%	27,43%	1,11%
(n=454)	fehlend	1,8%	1,8%	1,8%	93,2%		8,4%	0,0%	0,4%	0,7%
4	Anteil	19,37%	7,21%	5,41%	48,21%	21,43%	6,42%	3,14%	34,98%	4,63%
(n=223)	fehlend	0,4%	0,4%	0,4%	74,9%		2,2%	0,0%	0,0%	3,1%
5	Anteil	15,25%	1,98%	1,78%	67,65%	25,37%	3,10%	4,49%	24,51%	7,22%
(n=560)	fehlend	9,8%	9,8%	9,8%	51,4%		7,9%	0,5%	0,2%	3,6%
6	Anteil	20,56%	5,03%	9,44%	36,00%	10,00%	5,70%	3,02%	24,88%	4,62%
(n=209)	fehlend	13,9%	14,4%	13,9%	76,1%		7,7%	4,8%	1,9%	6,7%
7	Anteil	16,76%	11,89%	6,49%	75,00%	8,33%	11,23%	2,58%	30,50%	5,20%
(n=200)	fehlend	7,5%	7,5%	7,5%	94,0%		6,5%	3,0%	0,0%	13,5%
8	Anteil	14,93%	9,55%	4,18%	71,43%	36,73%	9,35%	3,55%	36,69%	4,17%
(n=338)	fehlend	0,9%	0,9%	0,9%	85,5%		5,0%	0,0%	0,0%	0,6%
9	Anteil	15,70%	6,25%	3,72%	48,48%	21,21%	6,36%	7,58%	23,11%	7,03%
(n=267)	fehlend	9,4%	10,1%	9,4%	87,6%		11,6%	1,1%	1,1%	4,1%
10	Anteil	24,68%	12,99%	4,76%	54,84%	16,13%	7,14%	3,98%	21,89%	5,15%
(n=233)	fehlend	0,9%	0,9%	0,9%	86,7%		3,9%	3,0%	0,0%	0,0%

<sup>1</sup> p-Wert auf Signifikanz (einzeln geprüft)

<sup>2</sup> p-Wert auf Unterschiede zwischen den Kliniken ( $\chi^2$ -Test)

tungsvergleich – ist es aber andererseits von größerem Interesse, inwieweit die **erwarteten Sterblichkeitsraten** durch eine Modellerweiterung noch variieren. Diese Raten können zunächst im marginalen Modell – d.h. ohne Einbeziehung der Klinikeffekte – anhand der Parameterschätzer im Gesamtmodell bestimmt werden. Die Hinzunahme von weiteren Variablen soll dann gestoppt werden, wenn in einem Erweiterungsschritt die Zunahme der Modellgüte gering ist oder wenn die Zahl der nicht mehr einschließbaren Fälle zu groß wird.

## 4.5 Modellaufbau zur Risikoadjustierung

### 4.5.1 Parametrischer Ansatz (logistische Regressionsanalyse)

Bei der Wahl der Reihenfolge der ins Modell einzuschließenden Variablen sind vielfältige Möglichkeiten denkbar. Bei den klassischen Selektionsverfahren (vgl. Kapitel 3.2.2) besteht die Schwierigkeit, dass vorab nur die Patienten einbezogen werden, bei denen Angaben für alle in Frage kommenden Parameter vollständig sind. Dies ist in der vorliegenden Fragestellung aber gerade unerwünscht, da die Zahl der letztlich verwendbaren Fälle möglichst hoch sein soll. Ein möglicher Ausweg stellt die Betrachtung fehlender Werte als zulässige Ausprägung dar. Wie im Exkurs über die demographischen Variablen erläutert wird, ist die Ausprägung „fehlend“ jedoch stark mit der Zielgröße (Tod) und mit der Dokumentationsqualität des Klinikums assoziiert. Hier soll die Reihenfolge absteigend nach der Vollständigkeit der Erfassung erfolgen.

Nach dem bis hierhin Diskutierten ergibt sich nun die folgende Anordnung der Modellbildung. Diese findet zunächst ohne Aufnahme des Klinikeffekts in das finale Modell statt.

0. keine Einflussgröße (nur Absolutglied);
1. Alter und Geschlecht;
2. pulmonaler Stau;
3. Rauchen, Hypercholesterinämie, Arterielle Hypertonie und Diabetes Mellitus;
4. Reanimationspflicht;
5. kardiogener Schock;
6. Zustand nach früherem Infarkt (Z.n.I), Manifeste Herzinsuffizienz (MHI), Niereninsuffizienz (NI).

Bei den in den Auswahlritten zusammengefassten Variablen sind die Angaben entweder bei (nahezu) allen Patienten verfügbar oder fehlend. Daher können diese Parameter blockweise ins Modell eingefügt werden.

**Tabelle 4.24: Modellbildung zur Risikoadjustierung (feste Effekte ohne Klinik)**

Schritt	vollst. [%]	Letalität [%]	Konkordanz [%]	erwartete Letalität [%] nach Klinik										
				1	2	3	4	5	6	7	8	9	10	
0	100,0	11,46	—	11,46	11,46	11,46	11,46	11,46	11,46	11,46	11,46	11,46	11,46	11,46
1	100,0	11,46	72,0	8,85	10,62	12,02	13,02	9,87	11,97	14,89	13,39	12,26	11,66	
2	99,4	11,39	76,5	7,62	10,65	11,85	13,88	9,79	11,87	15,06	14,50	11,59	11,03	
3	97,9	10,94	77,5	7,16	10,02	11,77	12,81	9,41	10,58	14,43	14,52	11,17	10,92	
4	96,1	10,93	81,0	6,95	9,79	11,51	12,52	9,60	10,06	14,36	14,20	12,59	11,26	
5	91,5	11,16	82,0	7,11	10,36	10,92	12,84	9,94	10,40	15,05	14,15	13,17	11,70	
6	86,5	11,24	82,5	7,42	10,50	10,96	12,90	9,50	10,97	15,90	14,08	12,87	12,14	

Die Tabelle 4.24 zeigt die Entwicklung der im vorigen Abschnitt beschriebenen Kriterien für jeden Schritt der Modellbildung aus obiger Aufstellung.

Durch Betrachtung des Anstiegs der Modellgüte (Anteil der konkordanten Paare, vgl. Kapitel 2.2.4) bei gleichzeitiger Abnahme der Vollständigkeit erscheint ein sinnvoller Stopp der Modellerweiterung nach Schritt 4 erreicht zu sein. An dieser Stelle ist die Vollständigkeit (insgesamt bei 96,1%) noch für jedes Klinikum bei mehr als 90% gegeben. Allerdings werden die Klinik-spezifischen Letalitätsraten der im Modell verbliebenen Daten unterschiedlich stark verändert. So sinkt die beobachtete Sterberate bei Klinik Nr. 6 von etwa 12,0% im vollständigen Modell (Schritt 1) auf 9,6% nach Schritt 4 (bei einer Datenvollständigkeit von 90%). Bei allen anderen Kliniken bleibt diese – bei deutlich höherer Datenvollständigkeit – nahezu unverändert. Bei Klinik Nr. 6 sind also bei relativ vielen verstorbenen Patienten die Risikofaktoren der Schritte 1 bis 4 lückenhafter erfasst als im Durchschnitt, wodurch die erwartete Sterblichkeit in dieser Klinik möglicherweise überschätzt wird, obwohl die erwartete Sterblichkeit bei Klinik Nr. 6 im Vergleich zu Schritt 1 stärker absinkt als bei den übrigen Häusern.

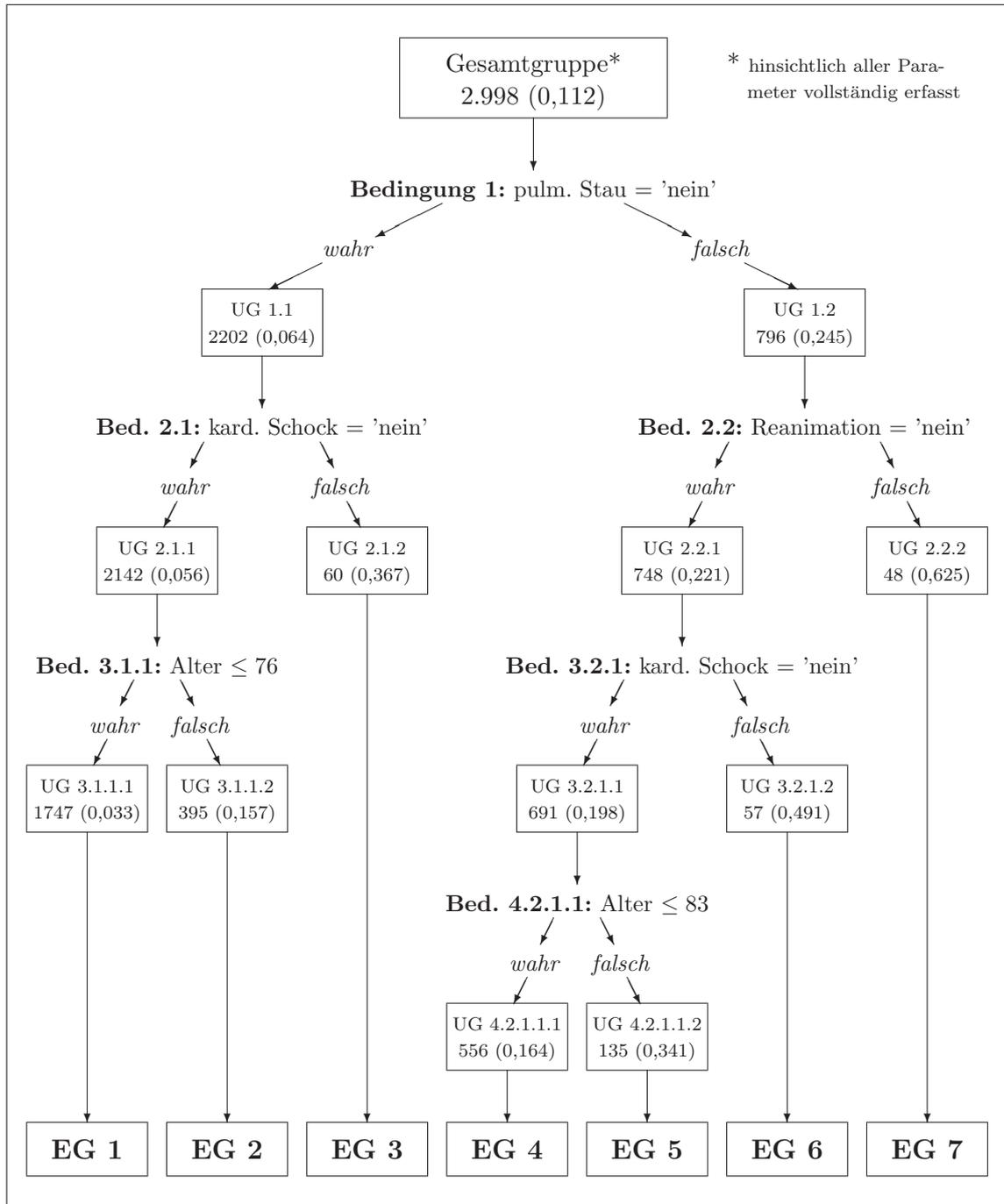
Das Modell Nr. 4 wird nun als finales Modell zur Risikoadjustierung betrachtet werden. Allerdings sollten die Ergebnisse, insbesondere wegen der erhöhten Rate fehlender Daten bei Klinik Nr. 6, denen des letzten vollständigen Modells (Schritt 1) gegenübergestellt werden. Die in diesem Ansatz nicht mehr betrachtete Variable „kardiogener Schock“ besitzt zwar – einzeln betrachtet – einen sehr starken Einfluss auf die Zielgröße; allerdings ist sie stark mit der bereits in Schritt 2 eingeschlossenen Variablen assoziiert.

### 4.5.2 CART-Analyse zur Exploration geeigneter Risikofaktoren

Alternativ zu der Strategie des Modellaufbaus aus den obigen Selektionsschritten ist es denkbar, die relevanten Risikofaktoren mittels eines Klassifikationsbaums zu ermitteln. Das Vorgehen bei der CART-Analyse im Exkurs zu den demographischen Variablen (Abschnitt 4.4.1) wird somit an dieser Stelle für die durch das Expertengremium vorgeschlagenen Parameter wiederholt (die Einstellungen sind hierbei wieder die gleichen wie im zitierten Abschnitt).

Zunächst wird das hinsichtlich der ausgewählten Parameter aus obiger Aufstellung bis einschließlich Modellschritt 6 vollständige Modell verwendet. Abbildung 4.9 zeigt die vom Algorithmus bestimmten Knoten und Endgruppen des vollständigen Modells.

Abbildung 4.9: CART-Analyse für mögliche Risikofaktoren, volles Modell



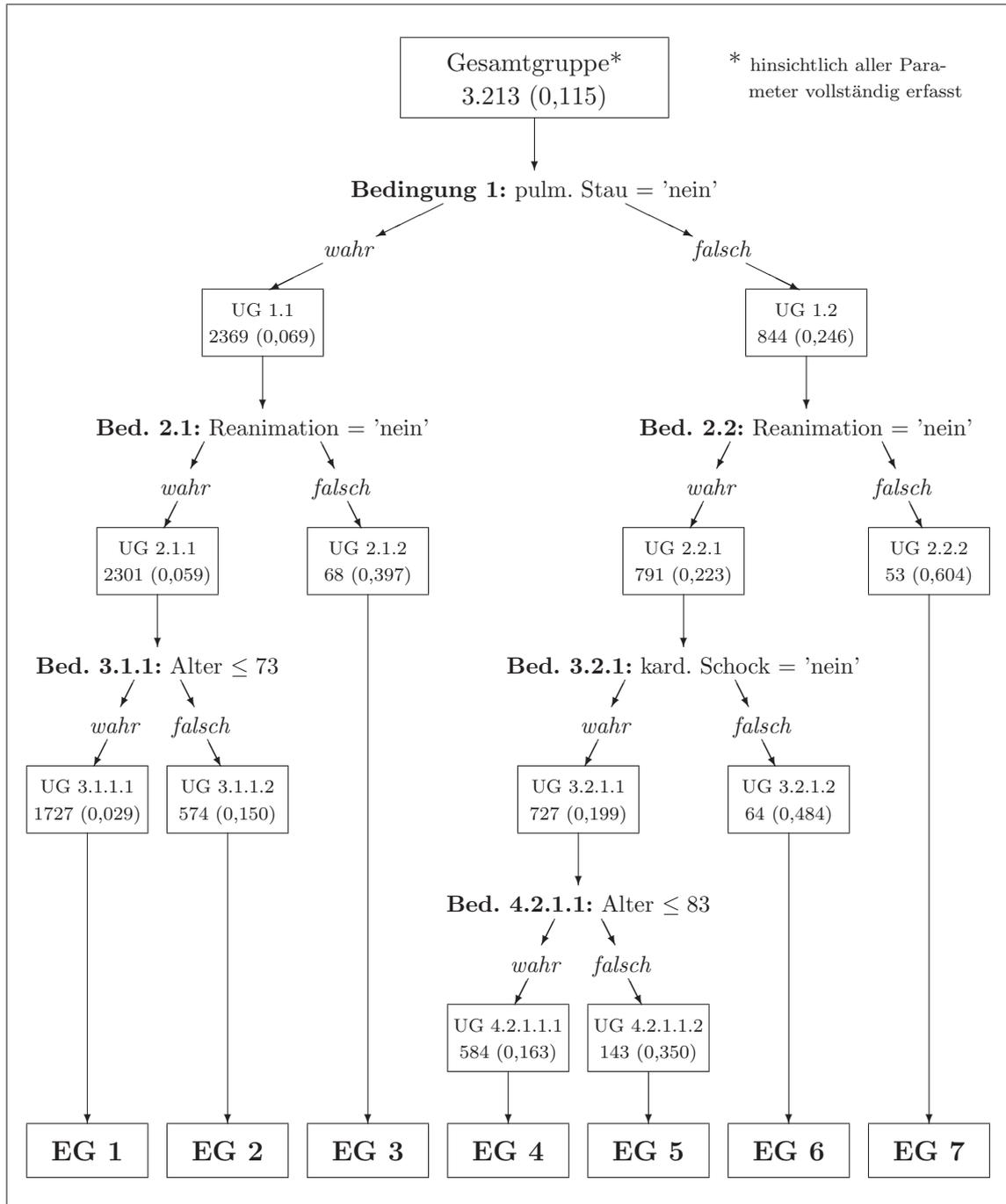
Unter den 12 Parametern aus Modell 6 wurden durch die CART-Analyse die Variablen pulmonaler Stau, kardiogener Schock, Reanimationspflicht und Alter für die Splits verwendet. Im Gegensatz zu den ausschließlich demographischen Variablen (vgl. Exkurs, Abschnitt 4.4.1) ist hier jedoch die Präsenz eines pulmonalen Staus das wichtigste Kriterium für die Wahrscheinlichkeit des Versterbens, und nicht mehr das Alter. Die Ergebnisse sind mit denen der entsprechenden logistischen Regressionsanalyse recht gut vergleichbar, da dort die drei erstgenannten Variablen unter den dichotomen Einflussgrößen die stärksten Odds Ratios bei jeweils hoher Signifikanz ( $p < 0,001$ ) zeigten.

Die 7 gebildeten Endgruppen weisen einerseits stark unterschiedliche Fallzahlen auf (weshalb auch die Tiefe der Split-Ebenen variiert), andererseits auch stark variierende Letalitätsraten. So zeigt Endgruppe 7 – d.h. reanimationspflichtige Patienten, bei denen ein pulmonaler Stau festgestellt wurde – mit 62,5% die mit großem Abstand höchste Sterblichkeitsrate aller Endgruppen. Die höchstens 76 Jahre alten Patienten, bei denen pulmonaler Stau und kardiogener Schock verneint wurde (Endgruppe 1) besitzen mit 3,3% Letalität die beste Prognose.

Wie bei der logistischen Regression sind auch bei der durchgeführten CART-Analyse nur Patienten zu berücksichtigen gewesen, bei denen alle Parameter des Modells Nr. 6 verfügbar sind (sonst würden fehlende Werte unerwünschterweise als 0 interpretiert).

In einem weitergehenden Schritt kann der CART-Algorithmus nun mit dem reduzierten Satz von 4 anstatt der 12 Parameter wiederholt werden, wodurch die Zahl der eingeschlossenen Patienten von 2.998 (oder 86,5% des Gesamtkollektivs) auf 3.213 (92,7%) erhöht wird. Die Ergebnisse des reduzierten Parametersatzes sind in Abbildung 4.10 dargestellt.

Abbildung 4.10: CART-Analyse für mögliche Risikofaktoren, reduziertes Modell



Es zeigen sich beim reduzierten Modell – bei prinzipiell gleicher Baumstruktur – an den Knoten 2.1 und 3.1.1 leicht veränderte Bedingungen im Vergleich zum vollen Modell. Das Kriterium „kardiogener Schock“ ist im reduzierten Modell nur noch für die nicht reanimationspflichtigen Patienten mit pulmonalem Stau (etwa 24,6% der Patienten) relevant. Im reduzierten Modell war diese Split-Variable für alle Endgruppen 1 bis 6 (98,4%) von Bedeutung.

## 4.6 Ergebnisse zum Einrichtungsvergleich

### 4.6.1 Hauptmodell, primäre Analyse im adjustierten GLMM

Als das nach den Modellerweiterungsschritten am ehesten geeignete Modell erscheint unter Abwägung der durch fehlende Werte gegebenen Problematik und der Vollständigkeit der Risikofaktoren das Modell nach Schritt Nr. 4. Dieses Modell enthält somit neben dem Faktor *Klinik*, welcher mit zufälligen Effekten in  $Z\gamma$  modelliert wird, die folgenden 8 Einflussgrößen mit festen Effekten in  $X\beta$ :

- $X_1$ : Alter;
- $X_2$ : Geschlecht;
- $X_3$ : pulmonaler Stau;
- $X_4$ : Rauchen;
- $X_5$ : Hypercholesterinämie (HCE);
- $X_6$ : Arterielle Hypertonie (AHT);
- $X_7$ : Diabetes Mellitus;
- $X_8$ : Reanimationspflicht;
- $Z_1$ : *Klinik*.

Das Modell, dessen Ergebnisse in Tabelle 4.25 dargestellt sind, hat somit die folgende Gestalt:

$$\text{logit}(p_{ij}) = X\beta + Z\gamma = \beta_0 + \beta_1 X_{1ij} + \dots + \beta_8 X_{8ij} + \gamma_i . \quad (4.2)$$

Der Parameter „Alter“ wird als stetige Größe aufgefasst; alle anderen festen Effekte sind von dichotomer Natur. Die dargestellten Odds Ratios werden für alle

dichotomen Variablen als Vorhandensein des Merkmals gegenüber der Abwesenheit dargestellt. Beim Alter bezeichnet das Odds Ratio die Zunahme des Mortalitätsrisikos für jeweils ein weiteres vollendetes Lebensjahr.

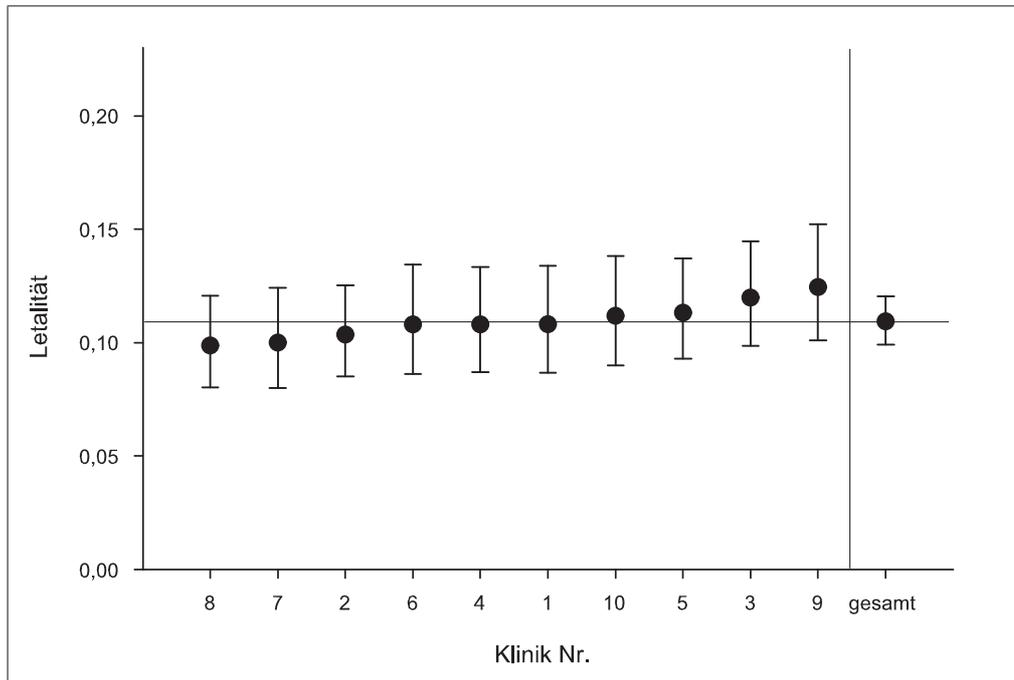
**Tabelle 4.25: Ergebnisse des risikoadjustierten Auswertungsmodells**  
– Schritt 4 (GLMM,  $n=3.330$ ,  $\tilde{p}=0,1093$ )

Parameter	OR	p-Wert	Mittelwert / Anteil nach Klinik									
			1	2	3	4	5	6	7	8	9	10
Alter	1,05	<0,001	62,94	65,70	67,15	68,45	63,46	65,10	70,10	69,25	67,31	65,71
Frauen	1,49	0,003	,2485	,3051	,3556	,3901	,3080	,3830	,4352	,4311	,3682	,4133
pulm. Stau	2,73	<0,001	,1394	,2695	,2756	,3498	,2412	,2181	,3161	,3623	,2326	,2267
Raucher	1,00	0,986	,4242	,3983	,3956	,3812	,4453	,3830	,3212	,2964	,3256	,4044
HCE	0,59	<0,001	,5667	,4814	,3333	,4709	,3432	,4521	,4093	,2904	,3915	,3689
AHT	0,91	0,470	,6121	,5949	,6667	,6054	,5362	,6330	,7668	,7066	,6667	,6178
Diab.Mell.	1,28	0,056	,2758	,2559	,2711	,2287	,1725	,2606	,3523	,3204	,2752	,3244
Reanimat.	11,29	<0,001	,0303	,0305	,0311	,0314	,0408	,0266	,0259	,0359	,0736	,0400
erwartete Letalität			,0695	,0979	,1151	,1252	,0960	,1006	,1436	,1420	,1259	,1126
beobachtete Letalität			,0667	,0881	,1356	,1211	,1020	,0957	,1088	,1138	,1667	,1200
OR (beobachtet/erwartet)			0,96	0,89	1,21	0,96	1,07	0,95	0,73	0,78	1,39	1,07
<i>Klinik</i> ( $\hat{\gamma}_i$ )		0,117*	-0,012	-0,061	0,104	-0,013	0,039	-0,014	-0,100	-0,114	0,147	0,025
<i>Klinik</i> ( $\hat{p}_i$ )			0,108	0,104	0,120	0,108	0,113	0,108	0,100	0,099	0,124	0,112
<i>p-Werte einzeln</i>			0,920	0,581	0,350	0,914	0,730	0,912	0,422	0,323	0,221	0,839
<i>Rangplatz</i>			6.	3.	9.	5.	8.	4.	2.	1.	10.	7.

\*  $\hat{\sigma}_a^2 = 0,0203$ ; SE = 0,0273; Z = 0,745

Als Gesamtergebnis kann aufgrund des p-Werts von  $p = 0,117$  für die globale Nullhypothese  $H_0 : \sigma_a^2 = 0$  **nicht** mit hinreichender Sicherheit auf Unterschiede zwischen Kliniken (im allgemeinen) – die in dieser Analyse betrachteten Kliniken sind als Elemente einer Zufallstichprobe vom Umfang 10 zu betrachten – geschlossen werden. Die dargestellten Effektschätzer sind somit lediglich als Zusatzinformation zu interpretieren. Selbst ohne Berücksichtigung dieses Testergebnisses zeigt keine Klinik einen zum Niveau  $\alpha = 5\%$  signifikant von 0 verschiedenen Effekt. Durch Betrachtung der Effektschätzer  $\hat{\gamma}_i$  belegt Klinik Nr. 8 in diesem Einrichtungsvergleich (Ranking) den Rangplatz 1 und Klinik Nr. 9 den letzten Platz, wenn die Effektschätzer aufsteigend sortiert werden.

**Abbildung 4.11: Krankenhaus-Letalität nach Klinikum – GLMM-basierte 95%-Vertrauensintervalle mit Adjustierung (Modell 4)**



Die Effektschätzer für die Einrichtung mit der geringsten beobachteten Sterberate (Klinik Nr. 1) liegt zwar nach der Adjustierung und Schrumpfung gerade noch im negativen Bereich (d.h. geringere Letalitätswahrscheinlichkeit als der Durchschnitt), sie belegt jedoch aufgrund des günstigeren Risikoprofils der eingeschlossenen Population lediglich einen mittleren Rangplatz. Bei Klinik Nr. 9 ist der (letzte) Rangplatz auch nach dieser Risikoadjustierung noch gegeben.

#### 4.6.2 Ergebnisse anderer Modellierungen / Robustheitsprüfung

Da die Auswahl der Adjustierungsparameter wie auch die Wahl der Modellklasse (hier: GLM oder GLMM) – wie zuvor vielfach besprochen – ausschlaggebend für das Ergebnis ist und somit durchaus strittig sein kann, seien abschließend Ergebnisse einiger anderer in Frage kommender Modellierungen dargestellt und diskutiert.

### 4.6.2.1 GLMM mit vollständiger Erfassung

Abgesehen vom Abgleich des Hauptmodells mit dem nach der CART-Analyse bestimmten Parametern, stellt sich die Frage, wie das Ergebnis ausgefallen wäre, wenn man anstatt nach Schritt 4 an einer anderen Stelle der Auswahlprozedur gestoppt hätte. Hierbei ist dasjenige Modell von Interesse, bei dem bei vollständiger Erfassung die meisten Parameter betrachtet werden können. In Tabelle 4.26 sind die Ergebnisse für das Modell nach Auswahl Schritt 1, d.h. mit lediglich dem Alter und dem Geschlecht als feste Einflussgrößen, dargestellt:

$$\text{logit}(p_{ij}) = X\beta + Z\gamma = \beta_0 + \beta_1 X_{1ij} + \beta_2 X_{2ij} + \gamma_i, \quad (4.3)$$

mit  $X_{1ij}$  : Alter und  $X_{2ij}$  : Geschlecht des  $j$ -ten Patienten in Klinik  $i$ .

**Tabelle 4.26: Ergebnisse des risikoadjustierten Auswertungsmodells  
– Schritt 1 GLMM,  $n=3.465$ ,  $\tilde{p}=0,1146$**

Parameter	OR	p-Wert	Mittelwert / Anteil nach Klinik									
			1	2	3	4	5	6	7	8	9	10
Alter	1,06	<0,001	62,89	65,46	67,24	68,45	63,59	66,22	70,42	69,14	67,40	65,67
Frauen	1,38	0,007	,2427	,3083	,3590	,3901	,3089	,3876	,4400	,4349	,3708	,4120
erwartete Letalität			,0885	,1062	,1202	,1302	,0987	,1197	,1489	,1339	,1226	,1166
beobachtete Letalität			,0819	,0892	,1366	,1211	,1107	,1196	,1150	,1154	,1685	,1245
OR (beobachtet/erwartet)			0,92	0,82	1,16	0,92	1,14	1,00	0,74	0,84	1,45	1,08
<i>Klinik</i> ( $\hat{\gamma}_i$ )		0,120*	-0,025	-0,092	0,068	-0,024	0,061	-0,000	-0,080	-0,064	0,135	0,021
<i>Klinik</i> ( $\hat{p}_i$ )			0,112	0,106	0,122	0,112	0,121	0,115	0,107	0,108	0,129	0,117
<i>p-Werte einzeln</i>			0,815	0,346	0,496	0,827	0,542	0,997	0,470	0,541	0,205	0,845
<i>Rangplatz</i>			4.	1.	9.	5.	8.	6.	2.	3.	10.	7.

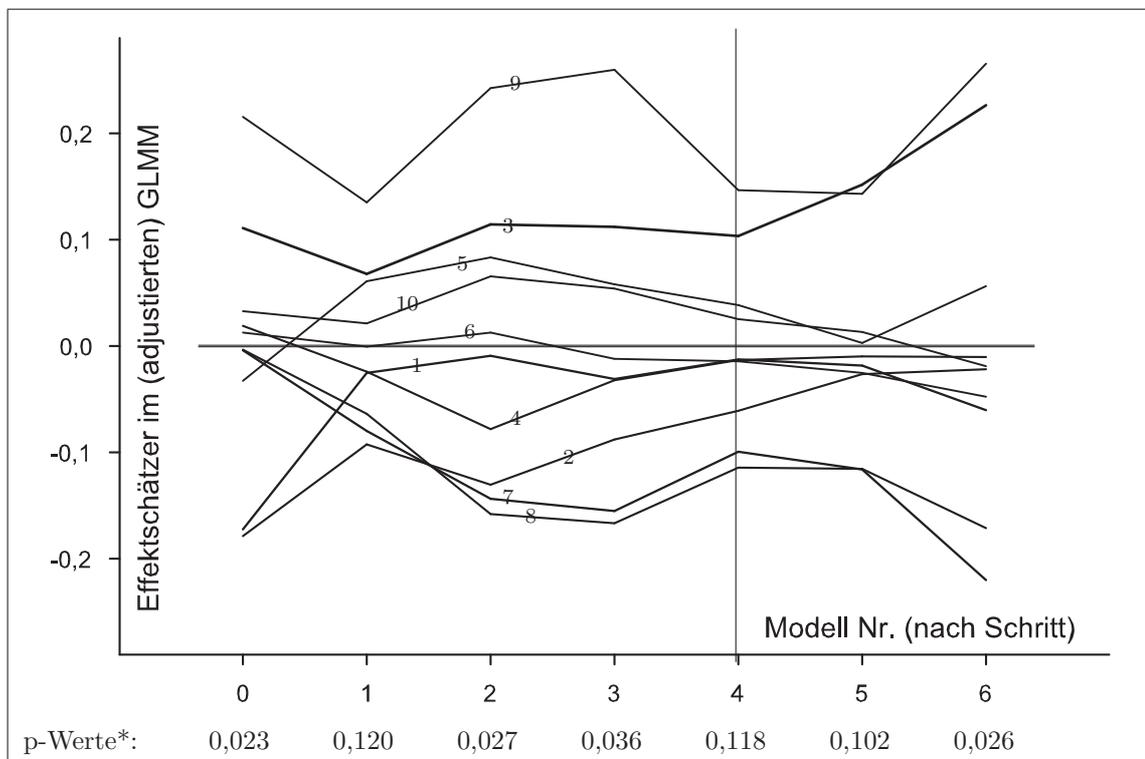
\*  $\hat{\sigma}_a^2 = 0,0159$ ; SE = 0,0216;  $Z = 0,739$

Im Vergleich zum Hauptmodell zeigen sich in diesem reduzierten, aber hinsichtlich der Beobachtungen vollständigen, Modell bezüglich der Beantwortung der Hauptfragestellung („Existieren Unterschiede zwischen Einrichtungen?“) praktisch keine Unterschiede. Zwar sind einige Wechsel in den Rangplätzen zu beobachten; abgesehen von Klinik 6 (die Klinik mit der höchsten Rate an Fehleingaben bei Modell 4) fallen diese jedoch eher gering aus.

#### 4.6.2.2 GLMM, andere Modelle

Zur weiteren Robustheitsprüfung sei schließlich die Frage erörtert, wie das Ergebnis ausgefallen wäre, wenn man die Auswahlprozedur anstatt nach Schritt 4 (bzw. Schritt 1) an einer anderen Stelle gestoppt oder eine hiervon abweichende Auswahl getroffen hätte. Hierzu sind in Abbildung 4.12 die Effektschätzer der Kliniken und die p-Werte der globalen Nullhypothese für jeden der Modellschritte dargestellt.

Abbildung 4.12: Effekt-Schätzer für Kliniken, GLMM adjustiert (Modelle 1 bis 6)



Es zeigt sich, dass die Testentscheidung (\*) für  $H_0 : \sigma_a^2 = 0$  bei dem gewählten Signifikanzniveau von  $\alpha = 0,05$  durchaus von der Stoppstelle der Modellerweiterung abhängt. Hätte man nach den Schritten 0, 2, 3 oder 6 abgebrochen, wäre die Entscheidung für  $H_1 : \sigma_a^2 > 0$  getroffen worden. Die Variation der  $\hat{\gamma}_i$  ist bei den Adjustierungen nach Modellschritt 1 (Alter und Geschlecht einbezogen) am geringsten. Nach Modellschritt 4 nimmt diese wieder zu, was zum einen in der Klinik-spezifisch verschiedenen Dokumentationsqualität begründet sein kann, andererseits aber auch auf tatsächlich unterschiedlichen Risikoprofilen in den Variablen der Schritte 5 und

6 basieren könnte. Diese Frage kann an dieser Stelle aufgrund der fehlenden Werte nicht endgültig geklärt werden.

Trotz der Ablehnung der globalen Nullhypothese bei einigen Modellen wäre in keinem Fall ein einzelner Klinikeffekt als signifikant verschieden von 0 eingestuft worden. Der kleinste Klinik-spezifische p-Wert wurde mit  $p = 0,064$  bei Modellschritt 3 für Klinik Nr. 9 beobachtet.

Von der Wahl des Adjustierungsmodells ist das Vorzeichen des Effektschätzers  $\hat{\gamma}_i$  bei den meisten teilnehmenden Einrichtungen weitgehend unabhängig. Sofern mindestens nach Alter und Geschlecht adjustiert wird, sind die Effektschätzer für die Kliniken Nr. 1, 2, 4, 7 und 8 stets negativ, und für Einrichtung Nr. 3, 5 und 9 stets positiv. Lediglich bei den übrigen beiden Häusern (Nr. 6 und 10) findet ein Wechsel des Vorzeichens statt. Insbesondere bei Klinik Nr. 6 kann wegen der, verglichen mit den übrigen Kliniken, deutlich erhöhten Rate fehlender Angaben (vgl. Tabellen 4.22 und 4.23) zumindest bezweifelt werden, ob die beobachtete Abnahme (Verbesserung) des Schätzers  $\hat{\gamma}_6$  bei den Modellen höherer Schrittnummer tatsächlich durch das in den entsprechenden Parametern überdurchschnittliche Risikoprofil begründet ist, oder ob die Fehleingaben hierfür ursächlich sind.

### **GLMM, Risikofaktoren nach CART-Analyse ausgewählt**

Für die Ergebnisse des Hauptmodells (4.2) kommt zusätzlich zu den Erweiterungsschritten ein Vergleich mit den laut CART-Analyse relevanten Parametern in Betracht. Die Adjustierungsparameter werden entsprechend der Split-Variablen der CART-Analyse – pulmonaler Stau, kardiogener Schock, Reanimationspflicht und Alter (vgl. Abbildungen 4.9 und 4.10) – verwendet.

Bei der Ergebnisinterpretation dieses Modells besteht die Schwierigkeit, dass die Fehleingabe-Rate für den Parameter „pulmonaler Schock“ zwischen den Kliniken sehr stark schwankt (vgl. Tabelle 4.23). Klinik Nr. 2 und Klinik Nr. 7 weisen bei dieser Variablen bei einer Gesamtrate von 5,2% über alle Kliniken Anteile von 14,7% bzw. 13,5% fehlender Werte auf. Durch diesen Umstand variiert entsprechend der Anteil der einschließbaren Fälle von lediglich 79,7% (Klinik 2) bis 99,4% (Klinik 8). Dennoch besitzt die Variable „Schock“ offensichtlich einen hohen Einflussgrad, wie sich beispielsweise an der erwarteten Letalitätsrate für Klinik Nr. 3, welche den geringsten Anteil von Patienten mit vorliegendem pulmonalen Schock aufweist, zeigt.

**Tabelle 4.27: Ergebnisse des risikoadjustierten Auswertungsmodells**  
 – Parameterauswahl nach CART (GLMM,  $n=3.213$ ,  $\tilde{p}=0,1093$ )

Parameter	OR	p-Wert	Mittelwert / Anteil nach Klinik									
			1	2	3	4	5	6	7	8	9	10
Alter	1,07	<0,001	62,89	65,67	67,13	68,45	63,45	65,57	70,60	69,15	67,37	65,77
pulm. Stau	2,64	<0,001	,1411	,2809	,2717	,3519	,2412	,2204	,3054	,3661	,2381	,2257
Reanimat.	7,81	<0,001	,0330	,0373	,0312	,0278	,0408	,0269	,0299	,0357	,0714	,0398
k. Schock	4,20	<0,001	,0300	,0432	,0111	,0463	,0724	,0484	,0479	,0417	,0714	,0487
erwartete Letalität			,0764	,1106	,1096	,1379	,1056	,1166	,1529	,1383	,1338	,1147
beobachtete Letalität			,0751	,1002	,1381	,1250	,1039	,1129	,1138	,1161	,1706	,1239
OR (beobachtet/erwartet)			0,98	0,90	1,30	0,89	0,98	0,96	0,71	0,82	1,33	1,09
<i>Klinik</i> ( $\hat{\gamma}_i$ )		0,109*	-0,004	-0,058	0,147	-0,039	-0,009	-0,010	-0,100	-0,091	0,131	0,032
<i>Klinik</i> ( $\hat{p}_i$ )			0,115	0,110	0,131	0,112	0,115	0,114	0,106	0,107	0,130	0,119
<i>p-Werte einzeln</i>			0,972	0,607	0,187	0,748	0,939	0,937	0,430	0,434	0,276	0,796
<i>Rangplatz</i>			7.	3.	10.	4.	6.	5.	1.	2.	9.	8.

\*  $\hat{\sigma}_a^2 = 0,0206$ ; SE = 0,0268;  $Z = 0,769$

Im Ergebnis belegt Klinik Nr. 9 – entgegen allen anderen in diesem Abschnitt dargestellten Modellen – nicht den 10. Rangplatz, der hier von Klinik Nr. 3 eingenommen wird.

Diesen „schlechtesten“ Rangplatz nimmt allerdings bei den Verhältnissen zwischen beobachteter und erwarteter Letalität (Odds Ratios) noch Klinik Nr. 9 ein, welche aber aufgrund der kleineren Fallzahl ( $n_9=267$  gegenüber  $n_3=454$ ) stärker zur Gesamtwahrscheinlichkeit bewegt („geschrumpft“) wird.

#### 4.6.2.3 Ein klassisches Generalisiertes Lineares Modell

Bei den bisher dargestellten Modellierungen wurden die teilnehmenden und in die Analyse aufgenommenen Notfallkliniken als Repräsentanten einer Zufallsstichprobe aus theoretisch beliebig vielen Kliniken aufgefasst. Die durch diese Annahme implizierte Aufblähung der beobachteten Unterschiede zwischen den Kliniken (vgl. Kapitel 2.4.2) wurde durch die hierarchische Modellierung der Kliniken als zufällige Effekte mittels der (E)BLUP-Strategie und der damit verbundenen Schrumpfung der Variation der Effektschätzer berücksichtigt.

Würde man die Registerdatenbank des BHiR – entgegen der dieser Arbeit zugrunde liegenden Philosophie – als prospektive Studie auffassen oder würde man sich eher für Unterschiede zwischen *genau diesen* teilnehmenden Kliniken interessieren als für Unterschiede im Behandlungserfolg zwischen Kliniken im allgemeinen, so ist die Entscheidung, die Sterblichkeitswahrscheinlichkeit über ein gemischtes Modell zu modellieren, durchaus zu kritisieren. In diesem Fall könnte alternativ ein klassisches Generalisiertes Lineares Modell bevorzugt werden. Zum Vergleich seien an dieser Stelle die Ergebnisse des GLMM (Modell 4) denen eines klassischen GLM – d.h. die Einflüsse der Kliniken werden als feste Effekte modelliert – gegenübergestellt. Diese Gegenüberstellung ist jedoch nicht als Robustheitsprüfung des primären Analysemodells aufzufassen, da es sich bei der Frage, ob ein GLM oder ein GLMM angepasst wird, weniger um die Prüfung von Annahmen, sondern eher um eine grundsätzliche Überlegung zur Situation bei Registerdaten bzw. des vorliegenden Registers handelt.

**Tabelle 4.28: Ergebnisse des risikoadjustierten Auswertungsmodells**  
– Schritt 4 (GLM,  $n=3.330$ ,  $\tilde{p}=0,1093$ )

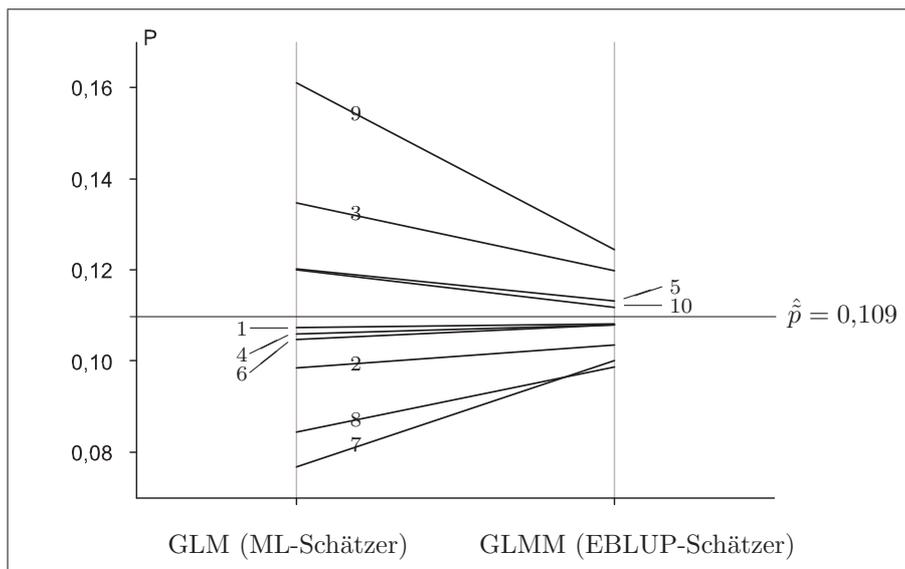
Parameter	OR*	p-Wert	Mittelwert / Anteil nach Klinik									
			1	2	3	4	5	6	7	8	9	10
Alter	1,05	<0,001	62,94	65,70	67,15	68,45	63,46	65,10	70,10	69,25	67,31	65,71
Frauen	1,49	0,003	,2485	,3051	,3556	,3901	,3080	,3830	,4352	,4311	,3682	,4133
pulm. Stau	2,73	<0,001	,1394	,2695	,2756	,3498	,2412	,2181	,3161	,3623	,2326	,2267
Raucher	1,00	0,986	,4242	,3983	,3956	,3812	,4453	,3830	,3212	,2964	,3256	,4044
HCE	0,59	<0,001	,5667	,4814	,3333	,4709	,3432	,4521	,4093	,2904	,3915	,3689
AHT	0,91	0,470	,6121	,5949	,6667	,6054	,5362	,6330	,7668	,7066	,6667	,6178
Diab.Mell.	1,28	0,056	,2758	,2559	,2711	,2287	,1725	,2606	,3523	,3204	,2752	,3244
Reanimat.	11,29	<0,001	,0303	,0305	,0311	,0314	,0408	,0266	,0259	,0359	,0736	,0400
erwartete Letalität*			,0695	,0979	,1151	,1252	,0960	,1006	,1436	,1420	,1259	,1126
beobachtete Letalität			,0667	,0881	,1356	,1211	,1020	,0957	,1088	,1138	,1667	,1200
OR (beobachtet/erwartet)			0,96	0,89	1,21	0,96	1,07	0,95	0,73	0,78	1,39	1,07
Klinik ( $\hat{\beta}_{9i}$ )		0,143	-0,022	-0,117	0,239	-0,036	0,107	-0,048	-0,389	-0,287	0,447	0,106
Klinik ( $\hat{p}_i$ )			0,107	0,098	0,135	0,106	0,120	0,105	0,077	0,084	0,161	0,120
p-Werte einzeln			0,923	0,452	0,114	0,867	0,497	0,847	0,106	0,112	0,015	0,623
Rangplatz			6.	3.	9.	5.	8.	4.	1.	2.	10.	7.

\* im GLM Modell ohne Klinik-Effekte

Verglichen mit den Ergebnissen im GLMM (Modell 4), zeigen sich beim GLM natürlich größere Unterschiede zwischen den Effekt-Schätzern. Aufgrund der fallzahlabhängigen Schrumpfung im GLMM wird Klinik Nr. 7 (mit  $n_7 = 193$  einschließbaren Fällen) im GLM als das „beste“ Haus eingestuft. Der Schrumpfungseffekt ist bei Klinik Nr. 8 beispielsweise wegen der höheren Fallzahl ( $n_8 = 334$ ) geringer, wodurch für GLMM Klinik Nr. 8 der beste Effekt geschätzt wird.

Auch wenn der p-Wert für den Einflussgrad des Faktors Klinik im GLM noch leicht über dem im GLMM liegt, würde der Effekt von Klinik Nr. 9, welches in beiden Modellen den Rangplatz 10 einnimmt, bei dieser Modellierung sogar als signifikant eingestuft werden. Die Schrumpfungswirkung der BLUP-Schätzer im GLMM auf der Originalskala ist für Modellschritt 4 in Abbildung 4.13 noch einmal illustriert.

**Abbildung 4.13: Adjustiertes Modell 4, Effekt-Schätzer (transformiert) im GLM und GLMM**



### 4.6.3 Fazit

Als Ergebnis dieser Auswertung kann festgehalten werden, dass aufgrund der vorliegenden Daten grundsätzlich nicht auf eine systematisch unterschiedliche Behandlungsqualität zwischen Kliniken im allgemeinen geschlossen werden kann. Zwar weisen die Kliniken hinsichtlich der beobachteten Letalitätsraten durchaus gewisse Un-

terschiede auf, jedoch sind diese zu einem großen Teil auf unterschiedliche Risiko-profile der Patientengruppen zurückzuführen.

Zur Adjustierung der Sterblichkeitsraten wurden mittels verschiedener Methoden im Register erhobene patientenbezogene Kovariaten, die als relevant bezüglich der Wahrscheinlichkeit, im Krankenhaus zu versterben, einzustufen sind, bestimmt und in die Analyse einbezogen. Bei der Auswahl dieser Parameter besteht aufgrund nicht gemachter Angaben die Schwierigkeit, die Vollständigkeit des Analysemodells (d.h. die Berücksichtigung möglichst vieler der in Betracht kommenden Kovariaten) gegen die hierzu im Spannungsverhältnis stehende Vollständigkeit der Fälle abzuwägen. Das Haupt-Analysemodell stellt somit einen Kompromiss aus diesen beiden Optimalitätskriterien dar.

Die Kliniken zeigen zumeist signifikante Unterschiede in der Verteilung der ausgewählten Risikofaktoren auf, wodurch die Letalitätsraten ohne eine geeignete Adjustierung nicht vergleichbar wären. Betrachtet man die Ergebnisse nach der Risiko-adjustierung in der gewählten Modellklasse, dem generalisierten *gemischten* linearen Modell, bei dem der Faktor *Klinik* mit zufälligen Effekten modelliert wurde, so ist die Variabilität der geschätzten Klinikeffekte um etwa zwei Drittel geringer als unter Verzicht auf diese Adjustierung. Die beobachteten Unterschiede sind somit zu einem großen Teil auf die unterschiedlichen Risikoprofile der Patientengruppen zurückzuführen. Da der Einflussfaktor *Klinik* in keiner der risikoadjustierten Analysen Signifikanz hinsichtlich der Zielgröße zeigte, müssen die verbliebenen Klinikeinflüsse als zufällig angesehen werden.

Für diesen Einrichtungsvergleich wurden die geschätzten Klinikeffekte in eine Rangfolge gebracht, bei der die Kliniken nach ihren Effektschätzern aufsteigend sortiert wurden. Das Klinikum mit dem kleinsten Schätzwert (und somit der kleinsten Krankenhaus-spezifischen Ereigniswahrscheinlichkeit) erhält somit den besten Rangplatz. Für diese Rangfolge ist es – trotz der Stabilität der Ergebnisse hinsichtlich der Gesamtinterpretation – durchaus von Bedeutung, nach *welchen* Kovariaten adjustiert wurde, da die geschätzten Effekte von vielen Kliniken nur geringfügig variieren.

---

## 5 Zusammenfassung und Diskussion

Die Überwachung der Versorgungsqualität im Gesundheitswesen besitzt in der anglo-amerikanischen Welt eine vergleichsweise lange Tradition, weil der Wettbewerb zwischen den Leistungserbringern in der medizinisch/ärztlichen Behandlung im Vergleich zu den meisten europäischen Ländern deutlich stärker ausgeprägt ist. In den letzten Jahren steigt das öffentliche Interesse an der Messung der Behandlungsqualität und deren Veröffentlichung – etwa anhand von Einrichtungsvergleichen – auch in Deutschland an.

Obwohl solche Vergleiche in einigen Einrichtungen in Deutschland bereits durchgeführt werden (und dies nicht nur im Bereich der öffentlichen Gesundheit), ist die Veröffentlichung der Ergebnisse insbesondere aus Sicht der medizinischen Einrichtungen selbst nach wie vor von hoher Brisanz. So ist etwa bei dem Anwendungsbeispiel dieser Arbeit – dem Berliner Herzinfarktregister – von allen teilnehmenden Einrichtungen zusätzlich zu ihrer grundsätzlichen Teilnahmebereitschaft noch einmal deren ausdrückliche Zustimmung zur Verwendung und Veröffentlichung der Daten und Ergebnisse einzuholen. Darüber hinaus ist die Identität jeder Klinik vor der Veröffentlichung der Ergebnisse so zu verblenden, dass in keinem Fall eine direkte Zuordnung zu den Kliniken herzustellen ist.

Diese allgemeine Skepsis gegenüber der Ergebnisveröffentlichung allerdings nur dann berechtigt, wenn methodisch nicht korrekte Vergleiche – beispielsweise begründet durch fehlende Berücksichtigung der Patientenprofile oder durch Wahl der falschen statistischen Modellklasse – Einrichtungen zu Unrecht im Vergleich zum allgemeinen Durchschnitt als besonders erfolgreich oder besonders erfolglos einstufen. Da diese Art von Vergleichsstudien in Deutschland noch wenig verbreitet ist, ist die Gefahr von unkorrekten Einrichtungsvergleichen meines Erachtens durchaus gegeben.

Ein häufig beobachteter Effekt von Einrichtungsvergleichen ist, dass durch wiederholt durchgeführte (z.B. jährliche) Analysen ein hohes Interesse in der Öffentlichkeit

und bei den teilnehmenden Zentren besteht. Dies führt dann meist dazu, dass sich die durchschnittliche Behandlungsqualität – zumindest bei genau dem betrachteten Untersuchungsgegenstand – im Laufe der Zeit kontinuierlich verbessert. Überspitzt formuliert kann das Wesen von Einrichtungsvergleichen so aufgefasst werden, dass sie selbst Ursache für (punktuell) verbesserte Behandlungsqualität sind. Eine so fokussierte Anstrengung birgt natürlich die Gefahr, dass die Versorgungsqualität in anderen – nicht gemessenen – Bereichen im Gegenzug nachlässt.

Mit dieser Arbeit soll ein Beitrag für die statistisch-methodischen Aspekte von Einrichtungsvergleichen geliefert werden. Es wurden statistische Methoden ausgewählt, die für die Analyse von Registerdaten und damit für den Vergleich der Versorgungsqualität von medizinischen Einrichtungen in Bezug auf die Beurteilung des Behandlungsergebnisses am Patienten geeignet sein können. Diese Behandlungsqualität kann durch einen einzigen oder auch durch mehrere Parameter gemessen werden. Im Falle einer einzelnen Zielgröße, hinsichtlich der ein Vergleich durchgeführt wird, spricht man häufig von „Rankings“; wenn mehrere Zielvariablen betrachtet werden, nennt man die Analyse auch „Profiling“. Je nach Art der interessierenden Zielgröße kann ein Ranking oder ein Profiling sinnvoll sein. Ist die Zielgröße, wie im Anwendungsbeispiel dieser Arbeit, das Überleben des Patienten (nach einem Herzinfarkt), so sind zusätzliche Zielvariablen weniger interessant als bei weniger lebensbedrohlichen Indikationen. Im Rahmen dieser Arbeit wurde der Terminus „Benchmarking“ als das Identifizieren einer besten Einrichtung mittels Erstellung einer Rangfolge hinsichtlich eines einzelnen Zielkriteriums verwendet.

Zur Analyse von Registrierungsdaten können lineare Modelle und generalisierte lineare Modelle angewandt werden. Da Krankenhäuser, die an einem Einrichtungsvergleich teilnehmen, zumeist nur eine Teilmenge aus allen denkbaren Krankenhäusern sind, ist es ein passender Ansatz, die Teilnehmer als Elemente einer (Zufalls-)Stichprobe zu betrachten. Demzufolge muss der Faktor *Klinik* hinsichtlich der Zielgröße als ein Faktor mit zufälligen Effekten modelliert werden. Wenn im Gegensatz hierzu Zentren als Faktoren mit festen Effekten modelliert würden, werden als Schätzung (adjustierte) Roh-Mittelwerte oder Anteile als Schätzungen verwendet. Somit würden beobachtete Unterschiede zwischen den Kliniken aufgrund von Effekten der Varianzaufblähung die tatsächlichen Klinikeffekte überschätzen. Als Folge könnten Kliniken zu Unrecht als „Sieger“ oder „Verlierer“ klassifiziert werden. Da der Ansatz der Roh-Mittelwerte zu irreführenden Interpretationen in der öffentlichen Wahrnehmung führen kann, sollten diese so nicht veröffentlicht werden.

Das in dieser Arbeit vorgeschlagene Konzept ist die Applikation von *gemischten* linearen Modellen, bei denen die Patienten als innerhalb der Kliniken unabhängige Messwiederholungen betrachtet werden. Diese Modellklasse wird auch als „Multilevel Models“ oder „hierarchische Modelle“ bezeichnet. Eine für die Einrichtungsvergleiche wesentliche Eigenschaft dieser Modelle ist die Verwendung der BESTEN LINEAREN UNVERFÄLSCHTEN PROGNOSE („best linear unbiased prediction [BLUP]), bei der die Variabilität der zufälligen Effekte ( $\rightarrow$  Kliniken) entsprechend des zu schätzenden Aufblähungsfaktors und der Stichprobengröße reduziert wird. Dieses Verfahren nennt man allgemein „Schrumpfung“ ( $\rightarrow$  shrinkage).

Registerdaten sind im Vergleich zu prospektiven (z.B. klinischen) Studien, bei deren Planung üblicherweise Annahmen getroffen werden, die letztendlich zu einer Fallzahl-Abschätzung führen, anders zu bewerten. Durch eine zufällige Zuweisung von Subjekten ( $\rightarrow$  Patienten) zu Behandlungen ( $\rightarrow$  Randomisierung) sollen systematische Effekte, z.B. die Beeinflussung des Behandlungsergebnisses durch andere Parameter, ausgeglichen werden. Solche Voraussetzungen sind in der Regel bei Registerdaten nicht gegeben. Vielmehr geschieht die Zuweisung der Patienten zu den Kliniken unkontrolliert, d.h. nicht durch einen Versuchsplan gesteuert. Da die Prognose der Patienten hinsichtlich des Behandlungserfolgs nicht nur von der Behandlung oder der Klinik, sondern auch von individuellen (z.B. demographischen oder anamnestischen) Eigenschaften abhängt und da diese (durchschnittliche) Prognose durch die unkontrollierte Situation zwischen den Einrichtungen schwanken kann, muss dies für die Bewertung des Behandlungserfolgs entsprechend berücksichtigt werden. Eine solche Berücksichtigung kann durch die Hinzunahme von prognostischen Parametern in das Analysemodell geschehen, was im Rahmen dieser Arbeit mit Hilfe der „Risikoadjustierung“ erfolgt.

Die Identifikation dieser Adjustierungsparameter kann mittels vielfältiger Verfahren geschehen. Ein nicht-methodischer Ansatz liegt etwa in der Auswahl durch Expertenbeurteilung. Steht diese Option nicht zur Verfügung, können zur Signifikanzprüfung klassische lineare Modelle (z.B. Modellselektionsmethoden) oder auch alternative Datenexplorations-Verfahren, wie beispielsweise Klassifikations- oder Regressionsbäume ( $\rightarrow$  CART-Methoden), zum Einsatz kommen. Für die Analyse des Beispieldatensatzes in dieser Arbeit wurde eine Kombination aus diesen drei Alternativen verwendet.

Bei hinreichend großen Datensätzen können theoretisch sehr viele Parameter zur

Adjustierung herangezogen werden, da die Problematik der Modell-Überspezifikation dann gering ist. Jedoch ergibt sich in der Praxis der Registerdaten häufig die Schwierigkeit, dass bei bestimmten Parametern eine hohe Rate von nicht gemachten Angaben vorliegt (so auch in der vorliegenden Datenbank). In diesem Fall sollte mit der Wahl der Modellparameter entsprechend sparsam umgegangen werden, damit nicht zu viele Fälle aufgrund unvollständiger Erfassung aus der Analyse ausgeschlossen werden. Dies ist insbesondere dann der Fall, wenn die Ausprägung „fehlend“ mit der Zielgröße assoziiert ist und wenn die Vollständigkeit zwischen den Einrichtungen schwankt, welches beides für das Anwendungsbeispiel gegeben ist.

Für die Zielgröße – die Krankenhaussterblichkeit nach Myokardinfarkt – wurden zur Adjustierung der Krankenhaus-spezifischen Letalitätsraten nach sorgfältiger Abwägung und Betrachtung mehrerer Auswahlmethoden die Parameter Alter, Geschlecht, Rauchen, pulmonale Stauung, Hypercholesterinämie, arterielle Hypertonie, Diabetes Mellitus und Reanimationspflicht verwendet. Das Vorhandensein des Merkmals erhöht bei allen binären Einflussgrößen – mit Ausnahme von Hypercholesterinämie und arterieller Hypertonie – die Letalitätswahrscheinlichkeit. In den beschriebenen Auswertungsmodellen sind mit Ausnahme des Alters alle beschriebenen Variablen binär. In dem zentralen Auswertungsmodell konnten nach dieser Auswahl 3.330 (96,1%) der initial 3.465 in der Datenbank vorhandenen Fälle einbezogen werden.

Die eingeschlossenen Populationen je Klinik weisen hinsichtlich dieser Adjustierungsvariablen signifikant verschiedene Verteilungen auf, so dass die Adjustierung als unbedingt notwendig anzusehen ist. Nach dieser Risikoadjustierung lässt das Ergebnis dieser Hauptanalyse nicht auf Unterschiede zwischen den teilnehmenden Einrichtungen schließen. Da die Auswahl der Modellparameter durchaus nicht eindeutig zu treffen ist, wurde die Analyse im Sinne einer Sensitivitätsanalyse (→ Robustheitsprüfung) für andere Auswahlalternativen wiederholt. Keines der alternativ angepassten Modelle zeigte ein in der Gesamtinterpretation verschiedenes Ergebnis. Lediglich die Rangplätze der Kliniken (d.h. Ordnung nach aufsteigender Sortierung der Einrichtungen nach ihrem individuellen Effekt-Schätzer) wiesen teils geringfügige Unterschiede auf. Aufgrund des in allen Analysen konsistenten Gesamtergebnisses „es kann nicht mit hinreichender Sicherheit auf Klinikunterschiede geschlossen werden“ kommt diesen Rangplätzen jedoch eine eher sekundäre Bedeutung zu.

Ein möglicher Grund für das Fehlen von signifikanten Unterschieden in der Be-

handlungsqualität könnte darin liegen, dass im Fall des akuten Myokardinfarkts bereits leitliniengerechte Therapien implementiert sind. Dadurch sind Unterschiede in der Behandlungsqualität kaum zu erwarten.

### **Ausblick**

In dieser Arbeit wurden empirische BLUP-Schätzer zur Bestimmung der Klinikeffekte in einem Generalisierten Gemischten Linearen Modell verwendet. Diese Schätzungen weisen die wünschenswerte Eigenschaft auf, dass sie den mittleren Prognosefehler über den gesamten Versuch hinweg minimieren. Jedoch ist die Variabilität dieser Schätzer geringer als die unbekannte tatsächliche Variation der Effekte. Zur Korrektur dieser konservativen Eigenschaft existieren für Gemischte Lineare Modelle weitergehende Vorschläge, wie beispielsweise der im Methodenteil angerissene „Triple Goal Schätzer“, der gewissermaßen einen Kompromiss zwischen den geschrumpften Schätzern und den Effektschätzern im klassischen linearen Modell (ML-Schätzer) darstellt. Da dieses Konzept jedoch bisher kaum bekannt und in keinem bekannten Softwarepaket verfügbar ist, kann es – insbesondere für generalisierte Modelle – kaum umgesetzt werden. Dieser Weg wurde daher nur am Rande diskutiert.

Zum Umgang mit fehlenden Werten existieren vielfältige Methoden und Ersetzungsansätze. Dieser Themenkomplex wurde ebenfalls nur am Rande diskutiert, beispielsweise mit dem Ergebnis, dass die in der vorliegenden Datenbank scheinbar zufällig fehlenden Angaben bei den Einflussgrößen keineswegs als zufällig fehlend verstanden werden können. Alternative Möglichkeiten, die Modelle durch geeignete Imputations-Mechanismen weiter zu optimieren, stellen für Einrichtungsvergleiche ein diskussionswürdiges Thema dar und sollten in weiterführenden Arbeiten behandelt werden.

Ein ebenfalls im Rahmen dieser Arbeit nur gestreiftes Thema ist die Berücksichtigung des Faktors „Zeit“ oder „Zeitpunkt“, etwa bei wiederholt durchgeführten Rankings. Im Rahmen dieser Arbeit wurden die Daten von fünf aufeinander folgenden Jahren für *einen* Einrichtungsvergleich gemeinsam betrachtet. Denkbar ist hier natürlich auch eine separate Betrachtung für mehrere Zeiträume, beispielsweise für jedes einzelne Jahr der Erhebung. Weiterhin könnte in eine Modellierung der Zeitpunkt (des Jahres) des Patienteneinschlusses als (fester) Einflussfaktor einbezogen werden. Mögliche Gründe für die durchaus vorhandenen Schwankungen in der Le-

talitätsrate *innerhalb* der einzelnen Kliniken *zwischen* den Jahren könnten an dieser Stelle ebenfalls untersucht werden.

Ein weiterer Aspekt, der im Rahmen dieser Arbeit kaum diskutiert wurde, ist dass es bei der Messung von Behandlungsqualität durchaus Faktoren geben kann, die nicht nur das Behandlungsergebnis sondern auch den Behandlungs*ansatz* selbst beeinflussen können. Diese wirken sich dann gewissermaßen indirekt auf das Behandlungsergebnis aus. So ist vorstellbar, dass in bestimmten Indikationen bestimmte (kostspielige) Behandlungen nur einer bestimmten Patientengruppe (z.B. jüngeren) zukommt. In diesem Zusammenhang, d.h. wenn eine Abhängigkeit oder Wechselwirkung zwischen Kovariaten und der Behandlung selbst besteht, spricht man auch von "confounding factors". Diese Faktoren erschweren durchaus den Vergleich der Behandlungsergebnisse, insbesondere wenn diese Patientengruppen von Einrichtung zu Einrichtung unterschiedlich definiert werden. Einer solchen Situation könnte bei der Analyse etwa durch geeignete Bildung von Untergruppen, Ausschluss von Patienten aus der Analyse oder Hinzunahme von Interaktionstermen in das statistische Modell Rechnung getragen werden.

Ein bekanntes Phänomen bei wiederholten Messungen ist die „Regression zum Mittelwert“ („regression towards the mean“). Kliniken, die zu einem bestimmten Zeitpunkt der Erhebung *zufällig* besonders extreme Werte aufweisen, zeigen oftmals in einer späteren Auswertung eher durchschnittliche Werte. Dies gilt natürlich insbesondere bei kleinen Fallzahlen oder bei besonderen, schwer zu modellierenden, externen Einflüssen.

---

# Literaturverzeichnis

- [1] Agresti, A. (1996):  
*An Introduction to Categorical Data Analysis.*  
John Wiley & Sons Inc., New York;  
232 Seiten.
- [2] AOK-Bundesverband, Forschungs- und Entwicklungsinstitut für das Sozial- und Gesundheitswesen Sachsen-Anhalt (FEISA), HELIOS Kliniken, Wissenschaftliches Institut der AOK (WIdO) (Hrsg.):  
*Qualitätssicherung der stationären Versorgung mit Routinedaten (QSR) – Abschlussbericht.*  
435 Seiten.
- [3] Bauer P., Hommel G., Sonnemann E. (1988):  
*Multiple Hypothesenprüfung / Multiple Hypotheses Testing.*  
Springer-Verlag GmbH.
- [4] Breiman L., Friedman J.H., Olshen R.A. and Stone C.J. (1984):  
*Classification and Regression Trees.*  
Wadsworth & Brooks/Cole, Monterey (California);  
358 Seiten.
- [5] Browne W.J., Subramanian S.V., Jones K. and Goldstein H. (2005):  
*Variance partitioning in multilevel logistic models that exhibit overdispersion.*  
Journal of the Royal Statistical Society / Series A (168);  
Seiten 599-613.

- [6] Bussche H. van den, Wegscheider K. and Zimmermann T. (2006):  
*Medizinische Fakultäten: Der Ausbildungserfolg im Vergleich (II)*.  
Deutsches Ärzteblatt (103), Ausgabe 34-35;  
Seiten A-2225/B-1925/ C-1861.
- [7] Cox D.R. and Snell E.J. (1989):  
*The Analysis of Binary Data*.  
Chapman and Hall, London;  
240 Seiten.
- [8] Dahms S. (2000):  
*Bestandsgesundheit und Lebensmittelsicherheit Beiträge der Biometrie und Epidemiologie*.  
Habilitationsschrift - Freie Universität Berlin;  
221 Seiten.
- [9] Dempster A.P., Rubin R.B. and Tstakawa R.K. (1981):  
*Estimation in Covariance Components Models*.  
Journal of the American Statistical Association (76);  
Seiten 341-353.
- [10] Endahl L.A. and Utzon J. (2002):  
*Ranking lists over hospital quality - do they serve as a guidance or are they misleading? On methodological problems in connection with announcement of quality indicators*.  
Ugeskr Laeger 164 (38);  
Seiten 4385-4388.  
*Artikel nur in Dänischer Sprache verfügbar; Deutsche Übersetzung durch Dr. Oke Gerke*
- [11] Everitt B.S. and Dunn G. (1998):  
*Statistical Analysis of Medical Data - New Developments*.  
Arnold, New York;  
340 Seiten.

- [12] Fahrmeir L. and Tutz G. (2001):  
*Multivariate Statistical Modelling Based on Generalized Linear Models.*  
Springer, New York / Berlin;  
548 Seiten.
- [13] Fisher W.D. (1958):  
*On Grouping for maximum heterogeneity.*  
Journal of the American Statistical Association (53);  
Seiten 789-798.
- [14] Goldstein H. and Spiegelhalter D.J. (1996):  
*League Tables and Their Limitations: Statistical Issues in Comparisons of Institutional Performance.*  
Journal of the Royal Statistical Society. Series A (Statistics in Society) (159  
No. 3);  
Seiten 385-443.
- [15] Gaylor D.W., Lucas, H.L. and Anderson, R.L. (1970):  
*Calculation of Expected Mean Squares by the Abbreviated Doolittle and Square  
Root Methods.*  
Biometrics (26));  
Seiten 641-655.
- [16] Hannan E.L., Kilburn H. Jr., O'Donnell J.F., Lukacik G. and Shields E.P.  
(1990):  
*Adult open heart surgery in New York State. An analysis of risk factors and  
hospital mortality rates.*  
Journal of the American Statistical Association (21);  
Seiten 2768-2774.
- [17] Hannan E.L., Kumar D., Racz M., Siu A.L. and Chassin MR (1994):  
*New York State's Cardiac Surgery Reporting System: four years later.*  
Annals of Thoracic Surgery (6);  
Seiten 1852-1857.

- [18] Hand D.J. (1997):  
*Construction and Assessment of Classification Rules.*  
John Wiley & Sons, Chichester;  
232 Seiten.
- [19] Hanley J.A. and McNeil B.J. (1982):  
*The Meaning and Use of the Area under a Receiver Operating Characteristic (ROC) Curve.*  
Radiology (143);  
Seiten 29-36.
- [20] Hartley H.O., Rao J.N.K. and LaMotte L. (1978):  
*A Simple Synthesis-Based Method of Variance Component Estimation.*  
Biometrics (34);  
Seiten 233-244.
- [21] Hartung J. (1981):  
*Non-negative Minimum Biased Invariant Estimation in Variance Component Models.*  
Annals of Statistics (9);  
Seiten 278-292.
- [22] Hartung J., Elpelt B., Klösener K.-H. (2005):  
*Statistik. Lehr- und Handbuch der angewandten Statistik, Auflage 14.*  
Oldenbourg;  
1004 Seiten.
- [23] Helmert F.R. (1876):  
*Diskussion der Beobachtungsfehler in Koppes Vermessung für die Gotthardtunnelachse.*  
Zeitschrift für das Vermessungswesen (V);  
Seiten 129-155.
- [24] Henderson C.R. (1984):  
*Applications of Linear Models in Animal Breeding.*  
University of Guelph.

- [25] Hill C.A., Winfrey K.L., Rudolph B.A. (1997):  
*“Best Hospitals“: A Description of the Methodology for the Index of Hospital Quality.*  
Inquiry 34 (1);  
Seiten 80-90.
- [26] Hocking R.R. (1976):  
*The Analysis and Selection of Variables in Linear Regression.*  
Biometrics (32);  
Seiten 1-50.
- [27] Howgill M., Blaza J., Cunningham L., Foster K.L. (2004):  
*The ratings game. How important are hospital rankings to consumers? Interview by Joyce Jensen.*  
Marketing Health Services (Spring 2004);  
Seiten 40-45.
- [28] Initiativkreis Ruhrgebiet (Hrsg.):  
*Klinikführer Rhein-Ruhr 2005/2006 (broschiert).*  
Klartext-Verlag;  
320 Seiten.
- [29] Judge G.G., Griffiths W.E., Hill R.C. and Lee T.C. (1980):  
*The Theory and Practice of Econometrics (Probability & Mathematical Statistics).*  
John Wiley & Sons Inc., New York;  
822 Seiten.
- [30] Kacker R.N. and Harville D.A. (1981):  
*Unbiasedness of two-stage estimation and prediction procedures for mixed linear models.*  
Communications in Statistics. Theory and Methods (10);  
Seiten 1249-1261.

- [31] Kackar R.N. and Harville D.A. (1984):  
*Approximations for Standard Errors of Estimators of Fixed and Random Effects in Mixed Linear Models.*  
Journal of the American Statistical Association (79);  
Seiten 853-862.
- [32] Kenward M.G. and Roger J.H. (1997):  
*Small Sample Inference for Fixed Effects from Restricted Maximum Likelihood.*  
Biometrics (53);  
Seiten 983-997.
- [33] Laird N.M. and Ware J.H. (1982):  
*Random-Effects Models for Longitudinal Data.*  
Biometrics (38);  
Seiten 963-974.
- [34] Leyland A.H. and Goldstein H. (2001):  
*Multilevel Modelling of Health Statistics.*  
John Wiley & Sons, Chichester;  
217 Seiten.
- [35] Littell R.C., Milliken G.A., Stroup W.W. and Wolfinger R.D. (1996):  
*SAS System for Mixed Models.*  
Cary, NC: SAS Institute, Inc.;  
656 Seiten.
- [36] Loyd J., Goldfield N. (2002):  
*The good, the bad, the ugly – Texas ranks hospital care.*  
Data Strategies & Benchmarks 6 (11);  
Seiten 172-174.
- [37] Maier B., Balzi D., Ainla T., Zeller M., Kallischnigg G., Barchielli A., Teesalu R., Cottin Y., Theres H., Buiatti E., Eha J. and Beer J.C. (2005):  
*Hospital Care of Patients with ST-Elevation Myocardial Infarction in Four Different European Regions.*  
Bundesgesundheitsblatt – Gesundheitsforschung – Gesundheitsschutz 9 (48);  
Seiten 1-8.

- [38] Maier B., Timme W., Kallischnigg G., Graf-Bothe C., Röhnisch J.U., Theres H. and Hegenbarth C. (2005):  
*Does Diabetes Mellitus Explain the Higher Hospital Mortality of Women with Acute Myocardial Infarction? Results from the Berlin Myocardial Infarction Registry.*  
Journal of Investigative Medicine 54 (3);  
Seiten 143-151.
- [39] McCullagh P. and Nelder J.A. (1989):  
*Generalized Linear Models, 2nd edition.*  
Chapman and Hall (London);  
320 Seiten.
- [40] Nightingale, F. (1863):  
*Notes on Hospitals (3rd edition).*  
London: Longman, Green, Longman, Roberts, and Green.
- [41] Nelder J.A. and Wedderburn R.W.M. (1972):  
*Generalized Linear Models.*  
Journal of the Royal Statistical Society, Series A (135);  
Seiten 761-768.
- [42] Pashova V. und Ulm K. (2000):  
*Two Survival Tree Models for Myocardial Infarction Patients.*  
Discussion papers: Institut für Medizinische Statistik und Epidemiologie, Technische Universität München;  
15 Seiten.
- [43] Patterson H.D. and Thompson R. (1971):  
*Recovery of Inter-Block Information when Block Sizes are Unequal.*  
Biometrika (58);  
Seiten 545-554.
- [44] Pfeiffer K.B., Pesec B. and Mischak R. (1994):  
*Stability of Regression Trees.*  
Computational Statistics 94 (in Dutter R. and Grossmann W.: *Proceedings in Computational Statistics, 11th Symposium Vienna*;  
Physica Verlag (Heidelberg).

- [45] Rao C.R. (1971):  
*Minimum Variance Quadratic Unbiased Estimation of Variance Components.*  
Journal of Multivariate Analysis (1);  
Seiten 445-456.
- [46] Rieve J.A. (2003):  
*Accountability and hospital rankings.*  
Case Manager (United States) 14 (2);  
Seite 29.
- [47] Robinson G.K. (1991):  
*That BLUP is a good thing: the estimation of random effects.*  
Statistical Science (6);  
Seiten 15-51.
- [48] Satterthwaite F.W. (1946):  
*An Approximate Distribution of Estimates of Variance Components.*  
Biometrics Bulletin (2);  
Seiten 110-114.
- [49] Scheffé H. (1959):  
*The Analysis of Variance.*  
John Wiley & Sons Inc.;  
476 Seiten.
- [50] Searle S.R., Casella G. and McCulloch C.E. (1992):  
*Variance Components.*  
John Wiley & Sons, New York;  
528 Seiten.
- [51] Self S.G. and Liang K.Y. (1987):  
*Asymptotic Properties of Maximum Likelihood Estimators and Likelihood Ratio Tests Under Nonstandard Conditions.*  
Journal of the American Statistical Association (82);  
Seiten 605-610.

- [52] Shen W. and Louis T. (1988):  
*Triple-goal estimates in two-stage hierarchical models.*  
Journal of the Royal Statistical Society (Series B) 60 (2);  
Seiten 455-471.
- [53] Swallow W.H. and Monahan J.F. (1984):  
*Monte Carlo Comparison of ANOVA, MIVQUE, REML, and ML Estimators of Variance Components.*  
Technometrics (28);  
Seiten 47-57.
- [54] Theres H., Maier B., Matteucci Gothe R., Schnippa S., Kallischnigg G., Schren K.P. and Thimme W. (2004):  
*Influence of Gender on Treatments and Short-Term Mortality of Patients With Acute Myocardial Infarction in Berlin.*  
Zeitschrift für Kardiologie (93);  
Seiten 954-963.
- [55] Ugolini C. and Nobilio L. (2004):  
*Risk Adjustment for Coronary Artery Bypass Graft Surgery: an Administrative Approach versus EuroSCORE.*  
International Journal for Quality in Health Care 16 (2);  
Seiten 157-164.
- [56] Venzon D.J. and Moolgavkar S.H. (1988):  
*A Method for Computing Profile-Likelihood Based Confidence Intervals.*  
Applied Statistics (37);  
Seiten 87-94.
- [57] Verbeke G. and Molenberghs G. (2000):  
*Linear Mixed Models for Longitudinal Data.*  
Springer-Verlag, New York / Berlin / Heidelberg;  
568 Seiten.

- [58] Wang O.J., Wang Y., Lichtman J.H., Bradley E.H., Normand S.T. and Krumholz H.M. (1987):  
*“America’s Best Hospitals“ in the Treatment of Acute Myocardial Infarction.*  
Archives of Internal Medicine 167 (13);  
Seiten 1345-1351.
- [59] Wegscheider K. (2004):  
*Methodische Anforderungen an Einrichtungsvergleiche („Profiling“) im Gesundheitswesen.*  
Zahnärztliche Fortbildung Qualität im Gesundheitswesen (98);  
Seiten 647-654.
- [60] Wernecke K.D., Possinger K., Kalb G. und Stein J. (1998):  
*Validating Classification Trees.*  
Biometrical Journal 40 (8);  
Seiten 993-1005.
- [61] Wolfinger R. and O’Connell M. (1993):  
*Generalized Linear Mixed Models: A Pseudo-Likelihood Approach.*  
Journal of Statistical Computation and Simulation (4);  
Seiten 233-243.

## Internet-Quellen

- [62] Information zur Rating-Methodik der AHRQ:  
[http://www.qualityindicators.ahrq.gov/psi\\_download.htm](http://www.qualityindicators.ahrq.gov/psi_download.htm)  
(letzter Zugriff: April 2007)
- [63] Internetseite des „Berliner Herzinfarktregister e.V.“ (BHiR):  
<http://www.berlinerherzinfarktregister.de>  
(letzter Zugriff: April 2007)
- [64] Verwendete Fragebögen des BHiR:  
<http://www.herzinfarktregister.de/fakten/untersuchung.htm>  
(letzter Zugriff: Februar 2007)

- [65] Wegscheider K. (2006):  
*Berliner Herzinfarktregister: Klinikvergleich 2004/5.*  
[http://www.herzinfarktregister.de/fakten/08.11.2006%20Klinikvergleich/Klinikvergleich\\_Daten%20BHIR\\_Wegscheider2%20Nov.%202006.pdf](http://www.herzinfarktregister.de/fakten/08.11.2006%20Klinikvergleich/Klinikvergleich_Daten%20BHIR_Wegscheider2%20Nov.%202006.pdf)  
(letzter Zugriff: März 2007)
- [66] Software-Diskussion des „Centre for Multilevel Modelling“ (Universität Bristol):  
<http://www.cmm.bristol.ac.uk/learning-training/multilevel-m-software/index.shtml>  
(letzter Zugriff: März 2007)
- [67] Internetseite der „Bundesgeschäftsstelle Qualitätssicherung“ (BQS):  
<http://www.bqs-online.de>  
(letzter Zugriff: April 2007)
- [68] Bumiller E. (1995):  
*Death-Rate Rankings Shake New York Cardiac Surgeons.*  
Aus dem Artikelarchiv der Zeitung „New York Times“  
<http://query.nytimes.com/gst/fullpage.html?sec=health&res=990CE1D6113DF935A3575AC0A963958260>  
(letzter Zugriff: April 2007)
- [69] Postleitzahlen der Städte und Gemeinden in Brandenburg:  
<http://www.plz-postleitzahl.com/de/index.cfm?parm=Brandenburg>  
(letzter Zugriff: April 2007)
- [70] Gesundheitsberichterstattung des Bundes / Robert-Koch-Institut:  
<http://www.gbe-bund.de>  
(letzter Zugriff: Juli 2005)
- [71] Internetseite des „Klinikführer Rhein-Ruhr“:  
<http://www.kliniken-rhein-ruhr.de>  
(letzter Zugriff: April 2007)
- [72] Statistisches Bundesamt (2005):  
*Pressemitteilung zur Todesursachenstatistik 2003:*  
<http://www.destatis.de/presse/deutsch/pm2005/p0600092.htm>  
(letzter Zugriff: April 2007)

- [73] Statistisches Bundesamt (2005):  
*Sterbefälle 2003 nach Todesursachen.*  
<http://www.destatis.de/basis/d/gesu/gesutab20.php>  
(letzter Zugriff: *Juli 2005*)
- [74] Statistisches Bundesamt (2006):  
*Kosten 2004 nach Krankheitsklassen und Alter in € je Einwohner der jeweiligen Altersgruppe.*  
<http://www.destatis.de/basis/d/gesu/gesutab23.php>  
(letzter Zugriff: *April 2007*)
- [75] Zeitung „Tagesspiegel“; Downloadbereich des „Klinikführer Berlin“:  
<http://www.tagesspiegel.de/kliniktest>  
(letzter Zugriff: *April 2007*)
- [76] Internetseite des THCIC (Texas):  
<http://www.dshs.state.tx.us/default.shtm>  
(letzter Zugriff: *April 2007*)
- [77] Zeitschrift „U.S. News & World Report“ zur Ranking-Methodologie:  
[http://www.usnews.com/usnews/health/best-hospitals/methodology\\_report.pdf](http://www.usnews.com/usnews/health/best-hospitals/methodology_report.pdf)  
(letzter Zugriff: *April 2007*)

---

## B Nebenrechnungen

### B.1 Rechnung zur Variation der binomialverteilten Zentrumshäufigkeiten

An dieser Stelle wird die auf Seite 97 im Kapitel 2.4.2.1 angesprochenen eigenen Überlegungen für die Variation der  $y_i$  ausgeführt.

Im Folgenden seien aus Gründen der Übersichtlichkeit die Varianzen  $Var(\mu)$  mit  $v$ ,  $Var(y_{(i)})$  mit  $v_{(i)}$  abgekürzt und  $(1 - p_i)$  mit  $q_i$  bezeichnet.

Vorüberlegung:

$$\begin{aligned} Var(y_i) &= E(y_i^2) - E(y_i)^2 = n_i p_i q_i \\ \iff E(y_i^2) &= v_i + \mu_i^2 = n_i p_i q_i + n_i^2 p_i^2 = n_i p_i (q_i + n_i p_i) = \mu_i (q_i + \mu_i) \end{aligned}$$

Somit gilt für

- $p = 2$ :

$$\begin{aligned} Var(y_1, y_2) &= E\left(\frac{1}{p-1} \sum_{i=1}^p (y_i - \bar{y})^2\right) \\ &= E((y_1 - \bar{y})^2 + (y_2 - \bar{y})^2) \\ &= E(y_1^2) - 2E(y_1 \bar{y}) + E(\bar{y}^2) + E(y_2^2) - 2E(y_2 \bar{y}) + E(\bar{y}^2) \\ &= E(y_1^2) - (E(y_1^2) + E(y_1)E(y_2)) + E(\bar{y}^2) \quad | \quad \bar{y} = \frac{y_1 + y_2}{2} \\ &\quad + E(y_2^2) - (E(y_2^2) + E(y_1)E(y_2)) + E(\bar{y}^2) \\ &= -2(E(y_1)E(y_2) - E(\bar{y}^2)) \\ &= -2(\mu_1 \mu_2 - v - \mu^2) = 2(v + \mu^2 - \mu_1 \mu_2) \end{aligned}$$

- $p = 3$ :

$$\begin{aligned}
 \text{Var}(y_1, y_2, y_3) &= E \left( \frac{1}{p-1} \sum_{i=1}^p (y_i - \bar{y})^2 \right) \\
 &= \frac{1}{2} E \left( (y_1 - \bar{y})^2 + (y_2 - \bar{y})^2 + (y_3 - \bar{y})^2 \right) \\
 &= \frac{1}{2} \left( v_1 + \mu_1^2 - \frac{2}{3} (v_1 + \mu_1^2 + \mu_1\mu_2 + \mu_1\mu_3) + v + \mu^2 \right. \\
 &\quad \left. + v_2 + \mu_2^2 - \frac{2}{3} (v_2 + \mu_2^2 + \mu_2\mu_1 + \mu_2\mu_3) + v + \mu^2 \right. \\
 &\quad \left. + v_3 + \mu_3^2 - \frac{2}{3} (v_3 + \mu_3^2 + \mu_3\mu_1 + \mu_3\mu_2) + v + \mu^2 \right) \\
 &= \frac{3-2}{3} \sum_{i=1}^3 (v_i + \mu_i^2) - \frac{2}{3} (2(\mu_1\mu_2 + \mu_1\mu_3 + \mu_2\mu_3) + 3(v + \mu^2)) \\
 &= -2(\mu_1\mu_2 + \mu_1\mu_3 + \mu_2\mu_3) + 3(v + \mu^2)
 \end{aligned}$$

- $p$  beliebig:

Wegen

$$p(v + \mu^2) = -1/2 \left( \sum_{i=1}^p \mu_i \left( \sum_{j \neq i} \mu_j \right) \right)$$

bzw.

$$E(y_i \bar{y}) = \frac{v_i + \mu_i p \mu}{p}$$

gilt

$$\begin{aligned}
\text{Var}(y) &= E \left( \frac{1}{p-1} \sum_{i=1}^p (y_i - \bar{y})^2 \right) \\
&= \frac{1}{p-1} \left( \sum v_i + \sum \mu_i^2 - 2 \sum \frac{v_i + p \mu_i \mu}{p} + \sum (v + \mu^2) \right) \\
&= \frac{1}{p-1} \left( \frac{p-2}{p} \sum v_i + \sum \mu_i^2 - 2 \sum \mu_i \mu + p(v + \mu^2) \right) \\
&= \frac{1}{p-1} \left( p(p-1)v + \sum \mu_i^2 - 2 \sum \mu_i \mu + p \mu^2 \right) \\
&= \frac{1}{p-1} \left( p(p-1)v + \sum \mu_i^2 - p \mu^2 \right) \\
&= \frac{1}{p-1} \left( p(p-1)v + p \mu^2 + (p-1) \text{Var}(E_i) - p \mu^2 \right) \\
&= \frac{1}{p-1} \left( p(p-1)v + (p-1) \text{Var}(E_i) \right) \\
&= p v + \text{Var}(E_i)
\end{aligned}$$

## B.2 Rechnung zum Impurity-Maß

An dieser Stelle wird die auf Seite 125 im Kapitel 2.6.3.1 angesprochenen eigenen Überlegungen für die drei dort eingeführten Impurity-Maße ausgeführt.

### Nebenrechnung 1 (Minima):

Zunächst wird eine binäre Zielgröße, also  $m = 2$ , betrachtet.

- Zur Bestimmung des Minimums für  $i_1(g)$  muss gelten:

$$\begin{aligned}
1 - \max_j \{\hat{p}(j|g)\} = \min! &\iff \hat{p}(j|g) = \max_j! \\
&\iff \hat{p}(j|g) = 1 \quad \text{für ein } j \in \{1; 2\}
\end{aligned}$$

$$\text{Also gilt:} \quad \exists j : \hat{p}(j|g) = 1 \quad \implies i_1(g) = 0$$

Aus dem Wertebereich von  $\hat{p}(j|g) \in [0; 1]$  folgt natürlich ebenfalls, dass  $i_1(g)$  keine Werte  $< 0$  annehmen kann, womit das Minimum in  $\hat{p}(j|g) \in \{0; 1\}$  festgestellt ist.

- Entsprechend gilt für  $i_2(g)$ :

$$2 \hat{p}(j|g) (1 - \hat{p}(j|g)) = \min! \iff \hat{p}(j|g) = 0$$

$$\iff \hat{p}(i|g) = 1 \quad \text{für } i \neq j$$

Also gilt:  $\exists j : \hat{p}(j|g) = 0 \implies i_2(g) = 0$

Da  $p(1-p)$  stets nicht-negativ für alle  $p \in [0; 1]$  ist, folgt, dass  $i_2(g)$  keine Werte  $< 0$  annehmen kann, womit gezeigt ist, dass das Minimum an der Stelle  $\hat{p}(j|g) \in \{0; 1\}$  liegt.

- Für  $i_3(g)$ :

Da  $p \in [0; 1]$  und wegen  $\log(p) \in [0; -1] \forall p \in [0; 1]$ ,  
gilt  $-p \log(p) \in [0; +1]$ .  
Gleiches gilt für  $(1-p) \in [0; +1]$ .

Der Term  $-p \log(p) - (1-p) \log(1-p)$  ist also stets  $\geq 0$ .

Sei nun ein  $\hat{p}(j|g) = 0, j \in \{1; 2\}$ ,

dann  $i_3(g) = -0 - 0 = 0$  da  $\log_b(1) = 0 \quad \forall b > 1$   
für  $\hat{p}(j|g) = 1$  analog.

Damit ist gezeigt, dass die Entropie in  $p = 1$  und  $p = 0$  den Wert 0 und damit das Minimum annimmt.

Nach gleicher Verfahrensweise kann gezeigt werden, dass diese Eigenschaften auch für nicht binäre kategorielle Zielgrößen gelten, also für  $m > 2$ , wenn ein  $\hat{p}(j|g) = 1$  und alle anderen  $\hat{p}(i|g) = 0, i \neq j$ , was an dieser Stelle jedoch nicht ausgeführt wird.

Alle drei Impurity-Maße sind demnach an den Rändern des  $m$ -dimensionalen Wahrscheinlichkeitsraums für  $p_j$  minimal und gleich 0. Als nächstes stellt sich die Frage nach den Maxima der Unreinheit.

### Nebenrechnung 2 (Maxima und Monotonität):

Zunächst wird wieder eine binäre Zielgröße, also  $m = 2$ , betrachtet.

- Zur Bestimmung des (einzigen) Maximums für  $i_1(g)$  muss gelten:

$$\begin{aligned} 1 - \max_j \hat{p}(j|g) = \max_j \hat{p}(j|g) &\iff \max_j \hat{p}(j|g) = \min_j \hat{p}(j|g) \\ &\iff \hat{p}(j|g) = \frac{1}{m} = \frac{1}{2} \quad \text{für } j = 1, 2 \end{aligned}$$

Um zu zeigen, dass das gefundene Maximum in  $p = 0,5$  tatsächlich eindeutig ist, wird die Monotonität von  $i_1(g)$  in den beiden Intervallen  $[0; 0,5[$  und  $]0,5; 1]$  benutzt:

$$\begin{aligned} \text{Sei o.B.d.A. } \hat{p}(1|g) < 0,5, \\ \text{dann gilt } i_1(g) &= 1 - \hat{p}(2|g) = \hat{p}(1|g). \end{aligned}$$

Für die erste Ableitung von  $i_1(g)$  nach  $\hat{p}(1|g)$  gilt trivialerweise

$$\begin{aligned} \frac{\partial i_1(g)}{\partial \hat{p}(1|g)} &= 1, \\ \text{und entsprechend für } \hat{p}(1|g) > 0,5: &\frac{\partial i_1(g)}{\partial \hat{p}(1|g)} = -1. \end{aligned}$$

Damit ist gezeigt, dass  $i_1(g)$  in  $\hat{p}(1|g)$  und damit spiegelbildlich in  $\hat{p}(2|g)$  eine Dreiecksfunktion darstellt.

- Zur Bestimmung des Maximums für  $i_2(g)$  bei  $m = 2$  kann die Differenzierbarkeit nach  $p$  im Intervall  $[0; 1]$  genutzt werden. Da

$$\begin{aligned} i_2(g) &= 2 \hat{p}(1|g) (1 - \hat{p}(1|g)) \\ &= 2 \hat{p}(1|g) - 2 \hat{p}(1|g)^2, \\ \text{gilt } \frac{\partial i_2(g)}{\partial \hat{p}(1|g)} &= 2 - 4 \hat{p}(1|g). \end{aligned}$$

Die Nullstellenbestimmung der Ableitung  $\frac{\partial i_2(g)}{\partial \hat{p}(1|g)}$  ergibt

$$\begin{aligned} 2 - 4 \hat{p}(1|g) &\stackrel{!}{=} 0 \\ \iff \hat{p}(1|g) &= \frac{1}{2}. \end{aligned}$$

Wegen der Eindeutigkeit dieser Lösung und der Positivität der zweiten Ableitung folgt, dass auch dieses Extremum ein eindeutig bestimmtes Maximum darstellt.

- Für die Maximierung von  $i_3(g)$  wird die Grundregel

$$\frac{\partial \log_b(p)}{\partial p} = \frac{1}{p \ln(b)}$$

genutzt. Bei  $m = 2$  gilt also

$$\frac{\partial i_3(g)}{\partial \hat{p}(1|g)} = -\frac{1}{\hat{p}(1|g) \ln(b)} + \frac{1}{(1 - \hat{p}(1|g)) \ln(b)}.$$

Die Nullstellenbestimmung der Ableitung  $\frac{\partial i_3(g)}{\partial \hat{p}(1|g)}$  ergibt

$$\begin{aligned} -\frac{1}{\hat{p}(1|g) \ln(b)} + \frac{1}{(1 - \hat{p}(1|g)) \ln(b)} &\stackrel{!}{=} 0 \\ \iff \frac{1}{\hat{p}(1|g)} &= \frac{1}{1 - \hat{p}(1|g)} \\ \iff \hat{p}(1|g) &= 1 - \hat{p}(1|g) . \end{aligned}$$

Diese Bedingung ist natürlich dann und nur dann erfüllt, falls

$$\hat{p}(1|g) = \frac{1}{2} .$$

Somit weisen alle drei Impurity-Maße die selben Extremstellen, nämlich bei  $p = \frac{1}{2}$ , auf. Es kann gezeigt werden, dass allgemein für  $m \geq 2$  das Maximum in  $\hat{p}(1|g) = \frac{i}{m}$  für alle  $i$  liegt. Dies wird hier jedoch nicht ausgeführt.

Schließlich ist noch zu betrachten, ob die Maße in  $p = \frac{1}{m}$  auch die selben Funktionswerte zeigen. Diese Forderung erscheint sinnvoll, da nur dann das Verhalten im gesamten Wahrscheinlichkeitsbereich vergleichbar ist.

### Nebenrechnung 3 (Normierung):

Zunächst wird wieder eine binäre Zielgröße, also  $m = 2$ , betrachtet.

- Bei  $i_1(g)$  liegt das Maximum in  $\hat{p}(1|g) = 0,5$  bei

$$i_1(g) = 1 - 0,5 = 0,5 .$$

- Genauso bei  $i_2(g)$ :  $i_2(g) = 2 * 0,5 - 2 * 0,25 = 0,5$  .

- Bei  $i_3(g)$  hängt der Funktionswert in  $\hat{p}(1|g) = 0,5$  von der Basis  $b$  des Logarithmus ab. Um alle drei Unreinheitsmaße in  $p = 0,5$  durch den selben Punkt – also durch

$i_3(g) = 0,5$  – zu „zwingen“, wird folgende Normierung durchgeführt.

Es muss also gelten:

$$i_3(g) = -\hat{p}(1|g) \log_b(\hat{p}(1|g)) - (1 - \hat{p}(1|g)) \log_b(1 - \hat{p}(1|g)) \stackrel{!}{=} 0,5$$

Und mit  $\hat{p}(1|g) = 1 - \hat{p}(1|g) = \frac{1}{2}$ :

$$-\hat{p}(1|g) \log_b(\hat{p}(1|g)) - (\hat{p}(1|g)) \log_b(\hat{p}(1|g)) \stackrel{!}{=} 0,5$$

$$\iff -2 * (\hat{p}(1|g)) \log_b(\hat{p}(1|g)) = 0,5$$

$$\iff (\hat{p}(1|g)) \log_b(\hat{p}(1|g)) = -1$$

Wegen  $\hat{p}(1|g) = \frac{1}{m} = 0,5$ :

$$\log_b(0,5) = -0,5$$

$$\iff b^{-0,5} = 0,5$$

$$\iff b = 4$$

Die gefunden Ergebnisse lassen sich leicht auf  $m > 2$  verallgemeinern.

Die Minimalstellen liegen natürlich dann vor, wenn ein  $j$  existiert, für das gilt:  $\hat{p}(j|g) = 1$ . In diesen Fällen nehmen alle Impurity-Maße den Wert 0 an.

Die (jeweils eindeutigen) Maximalstellen liegen, wie in Nebenrechnung 2 beschrieben, in

$$\hat{p}(1|g) = \frac{1}{m}$$

für alle  $i$ .

- Bei  $i_1(g)$  beträgt das Maximum immer

$$\begin{aligned} \max i_1(g) &= 1 - \max_j \hat{p}(j|g) \\ &= \frac{m-1}{m} . \end{aligned}$$

- Bei  $i_2(g)$  lässt sich das Maximum allgemein bestimmen als

$$\begin{aligned} \max i_2(g) &= 2 * \binom{m}{2} * \frac{1}{m^2} \\ &= 2 * \frac{m * (m-1)}{2} * \frac{1}{m^2} \\ &= \frac{m-1}{m} . \end{aligned}$$

- Bei  $i_3(g)$  gilt mit  $b = m^{m/(m-1)}$ :

$$\begin{aligned}\max i_2(g) &= -m * \left( \frac{1}{m} * \log_b \left( \frac{1}{m} \right) \right) \\ &= -\log_b \left( \frac{1}{m} \right) \\ &= \frac{m-1}{m} .\end{aligned}$$

**Bemerkung:**

Liegen gänzlich leere Zellen vor (d.h. existiert ein  $i$  mit  $\hat{p}(i|g) = 0, i \in \{1, \dots, m\}$ ), so ist die Entropie nicht definiert (Logarithmus strebt gegen  $-\infty$ ). Lässt man für  $\hat{p}(i|g) = \varepsilon (\varepsilon > 0)$  den Wert von  $\varepsilon$  gegen 0 streben, so strebt auch die Entropie gegen 0. Somit wird im Folgenden der Wert der Entropie in  $\hat{p}(i|g) = 0, i \in \{1, \dots, m\}$  als 0 angenommen.

---

# C Verwendete Analyseprogramme (Auszüge)

## C.1 SAS Programmcodes zur Simulation / LMM

### C.1.1 Programm zum Makro-Aufruf für Konfidenzintervalle (balancierter Fall, $n = 960$ , $p = 48$ , $\sigma_e^2 = 400$ , $\sigma_a^2 = 1600$ )

```
%LET dir = C:\...\Methoden\LMM;
%INCLUDE '&dir\prg\makro_konfidenz_LMM.sas';

libname simul '&dir\dat';

%LET truemean_total = %SYSEVALF(100)      ; *Gesamtmittel;
%LET s2_a             = %SYSEVALF(400)     ; *Streuung zwischen Kliniken;
%LET s2_e             = %SYSEVALF(1600)    ; *Streuung innerhalb der Kliniken;
%LET p                = %SYSEVALF(48)     ; *Anzahl Kliniken;
%LET balance          = 1                  ; *Für Daten setzen (0=unbalanciert, 1=balanciert);
%LET n_i              = %SYSEVALF(10)     ; *Fallzahl Klinik (balanciert);
%LET n_klein          = %SYSEVALF(0)      ; *Fallzahl kleine Kliniken (unbalanciert);
%LET n_gross          = %SYSEVALF(0)      ; *Fallzahl große Kliniken (inbalanciert);
%LET lauf             = %SYSEVALF(100000) ; *Anzahl Simulationsläufe;
%LET step             = %SYSEVALF(10000)  ; *Ausgabe auf Bildschirm bzw. Export nach;
%LET neu              = 1                  ; *Simulation fortsetzen (0=ja, 1=nein);

%gesamtmakro (dir=&dir, truemean_total=&truemean_total, s2_a=&s2_a, s2_e=&s2_e, p=&p, balance=&balance,
              n_i=&n_i, n_klein=&n_klein, n_gross=&n_gross, lauf=&lauf, step=&step, neu=&neu);
```

### C.1.2 Simulationsmakro

```
%macro daten;
  DATA WORK.a; s_a = SQRT(&s2_a); s_e = SQRT(&s2_e);
  CALL SYMPUT ('s_a',s_a); CALL SYMPUT ('s_e',s_e);
  RUN;
  DATA WORK.daten;
  DO klin = 1 TO &p;
    klin = klin; truemean_total = &truemean_total;
    IF &s_a EQ 0 THEN DO; truemean_klin = &truemean_total; END;
    ELSE DO truemean_klin = &truemean_total + rand('normal',0,&s_a); END;
  END;
%mend daten;
```

```

effect_klin = truemean_klin - truemean_total;
IF &balance EQ 1 THEN DO pat = 1 TO &n_i; y_ij = truemean_klin + rand('normal',0,&s_e); OUTPUT; END;
ELSE IF klin LE &p/2 THEN DO pat = 1 TO &n_klein; y_ij = truemean_klin + rand('normal',0,&s_e);
OUTPUT; END;
ELSE DO pat = 1 TO &n_gross; y_ij = truemean_klin + rand('normal',0,&s_e); OUTPUT; END;
END;
RUN;
%mend daten;

%macro auswerten;
PROC MEANS DATA=WORK.daten; VAR y_ij; BY klin; OUTPUT OUT=WORK.mittel mean=obsmean_klin; RUN;
PROC MIXED DATA=WORK.daten method=REML covtest;
CLASS klin; MODEL y_ij = / SOLUTION;
RANDOM klin / TYPE=vc SOLUTION CL; ODS OUTPUT covparms=WORK.cp solutionR=WORK.blup;
RUN; QUIT;
PROC TRANSPOSE DATA=WORK.cp (KEEP = covparm estimate) OUT=WORK.vc; RUN;
DATA WORK.vc; set WORK.vc; RENAME col1=hat_s2_a col2=hat_s2_e; lauf = &k; OUTPUT; RUN;
DATA WORK.cp (WHERE = (CovParm EQ 'klin')); SET WORK.cp;
IF probz GT 0.10 THEN DO test_s2a = 0; END;
IF probz LE 0.10 THEN DO test_s2a = 1; END;
IF probz EQ . THEN DO test_s2a = 0; END;
RUN;
DATA WORK.cp; SET WORK.cp; CALL SYMPUT ('testerg',test_s2a); RUN;
%LET h_1 = %SYSEVALF(&h_1 + &testerg);
%LET h_1t = %SYSEVALF(&h_1t + &testerg);
%LET power = %SYSEVALF(&h_1 / &k);
%mend auswerten;

%macro organisieren;
%IF &s_a GT 0 OR &testerg EQ 1 %THEN %DO;
DATA WORK.daten (WHERE = (pat EQ 1)); SET WORK.daten; RUN;
DATA WORK.daten (KEEP = klin truemean_klin); SET WORK.daten; RUN;
DATA WORK.zusammen; MERGE WORK.daten WORK.blup; BY klin;
klasse = int(truemean_klin/10 + 0.5) * 10; bereich = .;
RUN;
%END;
%mend organisieren;

%macro ergebnisse;
%IF &testerg EQ 1 %THEN %DO;
* Falls H_0 abgelehnt, Bereiche festlegen;
DATA WORK.klin_anh; SET WORK.zusammen;
lauf = &k; bereich = 0;
IF lower GT 0 THEN DO bereich = 1; END;
IF upper LT 0 THEN DO bereich = -1; END; OUTPUT;
RUN;
DATA WORK.klin_all; SET WORK.klin_all WORK.klin_anh; RUN;
%END;
%IF &s_a GT 0 %THEN %DO;
* Klassenhäufigkeiten ermitteln falls H1 gilt;
PROC FREQ DATA=WORK.zusammen; TABLES klasse / OUT=WORK.kl; RUN;
DATA WORK.kl (DROP = percent); SET WORK.kl; RENAME count=c_tmp; RUN;
DATA WORK.klassen (DROP = c_tmp count); MERGE WORK.klassen WORK.kl; BY klasse;
IF c_tmp EQ . THEN DO; c_tmp = 0; END; c_neu = count + c_tmp;
RUN;
DATA WORK.klassen; SET WORK.klassen; RENAME c_neu=count; RUN;
%END;
%mend ergebnisse;

%macro gesamtstats;
%IF &h_1t GT 0 %THEN %DO;
PROC SORT DATA=WORK.klin_all; BY klasse; RUN;
PROC FREQ DATA=WORK.klin_all; TABLES klasse*bereich / OUT=WORK.ks; RUN;
* Datei verkleinern;

```

```

DATA WORK.ks (DROP = percent); SET WORK.ks; RENAME count=c_tmp; RUN;
DATA SIMUL.ks (DROP = c_tmp count); MERGE SIMUL.ks WORK.ks; BY klasse bereich;
  IF c_tmp EQ . THEN DO; c_tmp = 0; END; c_neu = count + c_tmp;
RUN;
DATA SIMUL.ks; SET SIMUL.ks; RENAME c_neu=count; RUN;
PROC EXPORT DATA=SIMUL.ks OUTFILE='&dir\out\statest.xls' DBMS=EXCEL REPLACE; SHEET='dat'; RUN;
%IF &balance EQ 1 %THEN %DO;
  TITLE1 'sigma^2_a = &s2_a / sigma^2_e = &s2_e / p=&p / balanciert (q = &n_i)';
%END;
%ELSE %DO;
  TITLE1 'sigma^2_a = &s2_a / sigma^2_e = &s2_e / p=&p / unbalanciert (n_i = &n_klein / &n_gross)';
%END;
TITLE2 'Nach &k Läufen; power = &power';
PROC PRINT DATA=simul.ks LABEL NOOBS; RUN;
DATA WORK.klin_all; RUN;
%LET h_1t = %SYSEVALF(0);
%END;
%IF &s a GT 0 %THEN %DO;
* Klassenhäufigkeiten zusammenfassen falls H1 gilt;
DATA WORK.klassen; SET WORK.klassen; RENAME count=c_tmp; RUN;
DATA SIMUL.klassen (DROP = c_tmp count); MERGE SIMUL.klassen WORK.klassen; BY klasse;
  c_neu = count + c_tmp;
RUN;
DATA SIMUL.klassen; SET SIMUL.klassen; RENAME c_neu=count; RUN;
DATA WORK.klassen (DROP = c_tmp); SET WORK.klassen; count = 0; RUN;
PROC EXPORT DATA=SIMUL.klassen OUTFILE='&dir\out\classes.xls' DBMS=EXCEL REPLACE; SHEET='dat'; RUN;
%END;
DATA SIMUL.marke; letzter = &k; h_1 = &h_1; RUN;
%mend gesamtstats;

%macro gesammtmakro (truemean_total,s2_a=s2_e,p=,balance=,n_i=,n_klein=,n_gross=,lauf=,step=,neu=,dir=);
PROC DATASETS LIB=WORK MT=ALL KILL NOLIST; RUN; QUIT;
DATA WORK.klin_all; SET _NULL_; RUN;
DATA WORK.klassen; DO klasse = 0 TO 200 BY 10; count = 0; OUTPUT; END; RUN;
%LET h_1 = %SYSEVALF(0); %LET h_1t = %SYSEVALF(0); %LET power = %SYSEVALF(0);
%IF &neu EQ 1 %THEN %DO;
PROC DATASETS LIB=SIMUL MT=ALL KILL NOLIST; RUN; QUIT;
DATA SIMUL.ks; DO klasse=0 TO 200 BY 10; DO bereich = -1 to 1; count = 0; OUTPUT; END; END; RUN;
DATA SIMUL.klassen; DO klasse=0 TO 200 BY 10; count = 0; OUTPUT; END; RUN;
%LET start = 1;
%END;
%IF &neu EQ 0 %THEN %DO;
DATA SIMUL.marke; SET SIMUL.marke; CALL SYMPUT('letzter',letzter); CALL SYMPUT('h_1',h_1); RUN;
%LET start = %SYSEVALF(&letzter + 1);
%LET h_1 = %SYSEVALF(&h_1);
%END;
%LET halten = %SYSEVALF(&start + &step - 1);
%DO k = &start %TO &start + &lauf - 1;
PROC DATASETS LIB=WORK; SAVE klin_all klassen; RUN; QUIT;
%daten; %auswerten; %organisieren; %ergebnisse;
%IF &k EQ &halten %THEN %DO;
  DM 'clear log'; DM 'clear out';
  %gesamtstats; %LET halten = %SYSEVALF(&halten + &step);
%END;
%END;
%mend gesammtmakro;

```

## C.2 SAS Programmcodes zur Simulation / GLMM

### C.2.1 Programm zum Makro-Aufruf für Konfidenzintervalle (Situation d. BHiR, $\sigma_a^2 = 0,05$ )

```

%LET dir = C:\...\Methoden\LMM;
%INCLUDE '&dir\prg\makro_konfidenz_GLMM.sas';

libname simul '&dir\dat';

%LET p_total      = %SYSEVALF(397/3465) ; *Gesamtwahrscheinlichkeit;
%LET s2_a         = %SYSEVALF(0.05)    ; *Streuung zwischen Kliniken;
%LET p            = %SYSEVALF(10)      ; *Anzahl Kliniken;
%LET n_01        = %SYSEVALF(342)     ; *Fallzahl Klinik Nr. 1;
%LET n_02        = %SYSEVALF(639)     ; *Fallzahl Klinik Nr. 2;
%LET n_03        = %SYSEVALF(454)     ; *Fallzahl Klinik Nr. 3;
%LET n_04        = %SYSEVALF(223)     ; *Fallzahl Klinik Nr. 4;
%LET n_05        = %SYSEVALF(560)     ; *Fallzahl Klinik Nr. 5;
%LET n_06        = %SYSEVALF(209)     ; *Fallzahl Klinik Nr. 6;
%LET n_07        = %SYSEVALF(200)     ; *Fallzahl Klinik Nr. 7;
%LET n_08        = %SYSEVALF(338)     ; *Fallzahl Klinik Nr. 8;
%LET n_09        = %SYSEVALF(267)     ; *Fallzahl Klinik Nr. 9;
%LET n_10        = %SYSEVALF(233)     ; *Fallzahl Klinik Nr.10;
%LET lauf        = %SYSEVALF(100000)  ; *Anzahl Simulationsläufe;
%LET step        = %SYSEVALF(10000)   ; *Ausgabe auf Bildschirm bzw. Export nach;
%LET neu         = 1                  ; *Simulation fortsetzen (0=ja, 1=nein);

%gesamtmakro (dir=&dir, p_total=&p_total, s2_a=&s2_a, p=&p, n_01=&n_01, n_02=&n_02, n_03=&n_03,
              n_04=&n_04, n_05=&n_05, n_06=&n_06, n_07=&n_07, n_08=&n_08, n_09=&n_09, n_10=&n_10,
              lauf=&lauf, step=&step, neu=&neu);

```

### C.2.2 Simulationsmakro

```

%macro daten;
  DATA WORK.a; beta_0 = LOG(&p_total/(1 - &p_total)); s_a = SQRT(&s2_a);
  CALL SYMPUT ('beta_0',beta_0); CALL SYMPUT ('s_a',s_a);
  RUN;
  DATA WORK.daten;
  DO klin = 1 TO &p;
    klin = klin; p_total = &p_total; beta_0 = &beta_0;
    IF &s_a EQ 0 THEN DO; gamma_i = 0; END;
    ELSE DO; gamma_i = rand('normal',0,&s_a); END;
    p_i = 1 / (1 + exp(-beta_0 - gamma_i));
    IF klin EQ 1 THEN DO pat = 1 TO &n_01; tod = rand('bernoulli',p_i); OUTPUT; END;
    IF klin EQ 2 THEN DO pat = 1 TO &n_02; tod = rand('bernoulli',p_i); OUTPUT; END;
    IF klin EQ 3 THEN DO pat = 1 TO &n_03; tod = rand('bernoulli',p_i); OUTPUT; END;
    IF klin EQ 4 THEN DO pat = 1 TO &n_04; tod = rand('bernoulli',p_i); OUTPUT; END;
    IF klin EQ 5 THEN DO pat = 1 TO &n_05; tod = rand('bernoulli',p_i); OUTPUT; END;
    IF klin EQ 6 THEN DO pat = 1 TO &n_06; tod = rand('bernoulli',p_i); OUTPUT; END;
    IF klin EQ 7 THEN DO pat = 1 TO &n_07; tod = rand('bernoulli',p_i); OUTPUT; END;
    IF klin EQ 8 THEN DO pat = 1 TO &n_08; tod = rand('bernoulli',p_i); OUTPUT; END;
    IF klin EQ 9 THEN DO pat = 1 TO &n_09; tod = rand('bernoulli',p_i); OUTPUT; END;
    IF klin EQ10 THEN DO pat = 1 TO &n_10; tod = rand('bernoulli',p_i); OUTPUT; END;
  END;
%mend;

```

```

RUN;
%mend daten;

%macro auswerten;
PROC GLIMMIX DATA=WORK.daten;
  CLASS tod klin; MODEL tod(event='1')= / dist=binary link=logit SOLUTION;
  RANDOM klin / SOLUTION CL; NLOPTIONS ABSGCONV=0.002 1;
  ODS OUTPUT covparms=WORK.cp solutionR=WORK.blup;
RUN; QUIT;
DATA WORK.cp; SET WORK.cp; lauf = &k; check = 0;
  IF estimate NE . AND stderr NE . THEN DO; Z = estimate/stderr; END;
  IF estimate LE 0.0000001 THEN DO; Z = 0; END;
  IF estimate GT 0.0000001 AND stderr EQ . THEN DO; check = 1; END;
  test_s2a = 0; IF Z GT 0.95963281417 THEN DO; test_s2a = 1; END;
  CALL SYMPUT ('testerg',test_s2a); CALL SYMPUT ('est_s2a',estimate);
RUN;
%LET h_1 = %SYSEVALF(&h_1 + &testerg);
%LET h_1t = %SYSEVALF(&h_1t + &testerg);
%LET power = %SYSEVALF(&h_1 / &k);
%mend auswerten;

%macro organisieren;
%IF &s_a GT 0 OR &testerg = 1 %THEN %DO;
  DATA WORK.daten (WHERE = (pat EQ 1)); SET WORK.daten; RUN;
  DATA WORK.daten (KEEP = klin gamma_i); SET WORK.daten; RUN;
  DATA WORK.zusammen; MERGE WORK.daten WORK.blup; BY klin;
  klasse = int((gamma_i + 100) / 0.1 + 0.5) * 0.1 - 100; klasse = put(klasse,6.3); bereich = .;
RUN;
%END;
DATA WORK.blup (WHERE = (klin IN (2,7))); SET WORK.blup;
  lauf = &k; testerg = &testerg; IF stderrpred EQ . THEN DELETE;
RUN;
DATA WORK.blup_all; SET WORK.blup_all WORK.blup; RUN;
%mend organisieren;

%macro ergebnisse;
%IF &testerg EQ 1 %THEN %DO;
  * Falls H_0 abgelehnt, Bereiche festlegen;
  DATA WORK.klin_anh; SET WORK.zusammen; lauf = &k; bereich = 0;
  IF lower GT 0 THEN DO; bereich = 1;END;
  IF upper LT 0 THEN DO; bereich = -1;END; OUTPUT;
RUN;
  DATA WORK.klin_all; SET WORK.klin_all WORK.klin_anh RUN;
%END;
%IF &s_a GT 0 %THEN %DO;
  * Klassenhaeufigkeiten nach Kliniknummer hinzufügen falls H1 gilt;
  DATA WORK.klassen (KEEP = klin klasse count); MERGE WORK.klassen WORK.zusammen; BY klin klasse;
  IF gamma_i NE . THEN DO; count = count + 1; END;
RUN;
%END;
%mend ergebnisse;

%macro gesamtstats;
* Verteilung der Schätzer ausgeben fr 2 Kliniken;
%IF &k LE 27500 %THEN %DO;
  DATA SIMUL.blup_all; SET SIMUL.blup_all WORK.blup_all; RUN;
  PROC EXPORT DATA=SIMUL.blup_all OUTFILE='\&dir\out\blups.xls' DBMS=EXCEL REPLACE; SHEET='dat'; RUN;
%END;
DATA WORK.blup_all; SET _NULL_; RUN;
%IF &h_1t GT 0 %THEN %DO;
  PROC SORT DATA=WORK.klin_all; BY klin klasse; RUN;
  PROC FREQ DATA=WORK.klin_all; TABLES klasse*bereich / OUT=WORK.ks; BY klin; RUN;
  * Datei verkleinern;
  DATA WORK.ks (DROP = percent); SET WORK.ks; RENAME count=c_tmp; RUN;

```

```

DATA SIMUL.ks (DROP = c_tmp count); MERGE SIMUL.ks WORK.ks; BY klin klasse bereich;
  IF c_tmp EQ . THEN DO; c_tmp = 0; END; c_neu = count + c_tmp;
RUN;
DATA SIMUL.ks (WHERE = (klin NE .)); SET SIMUL.ks; RENAME c_neu=count; RUN;
PROC EXPORT DATA=SIMUL.ks OUTFILE='&dir\out\statest.xls' DBMS=EXCEL REPLACE; SHEET='dat'; RUN;
TITLE1 'sigma^2_a = &s2_a / nach &k Laufen; power = &power';
PROC PRINT DATA=simul.ks LABEL NOOBS; RUN;
DATA WORK.klin_all; RUN;
%LET h_1t = %SYSEVALF(0);
%END;
%IF &s_a GT 0 %THEN %DO;
* Klassenhufigkeiten zusammenfassen falls H1 gilt;
DATA WORK.klassen; SET WORK.klassen; RENAME count=c_tmp; RUN;
DATA SIMUL.klassen (DROP = c_tmp count); MERGE SIMUL.klassen WORK.klassen; BY klin klasse;
  c_neu = count + c_tmp;
RUN;
DATA SIMUL.klassen; SET SIMUL.klassen; RENAME c_neu=count; RUN;
DATA WORK.klassen (DROP=c_tmp); SET WORK.klassen; count = 0; RUN;
PROC EXPORT DATA=SIMUL.klassen OUTFILE='&dir\out\classes.xls' DBMS=EXCEL REPLACE; SHEET='dat'; RUN;
%END;
DATA SIMUL.marke; letzter = &k; h_1 = &h_1; RUN;
%mend gesamtstats;

%macro gesamtmakro
  (dir=,p_total=,s2_a=,p=,n_01=,n_02=,n_03=,n_04=,n_05=,n_06=,n_07=,n_08=,n_09=,n_10=,lauf=,step=,neu=);
PROC DATASETS LIB=WORK MT=ALL KILL NOLIST; RUN; QUIT;
DATA WORK.klin_all; SET _NULL_; RUN;
DATA WORK.blup_all; SET _NULL_; RUN;
DATA WORK.klassen; DO klin = 1 TO 10; DO klasse = -1.5 TO 1.5 BY 0.1; count = 0; klasse=put(klasse,6.3);
  OUTPUT; END; END;
RUN;
%LET h_1 = %sysevalf(0); %LET h_1t = %sysevalf(0); %LET power = %sysevalf(0);
%IF &neu EQ 1 %THEN %DO;
  PROC DATASETS LIB=SIMUL MT=ALL KILL NOLIST; RUN; QUIT;
  DATA SIMUL.blup_all; SET _NULL_; RUN;
  DATA SIMUL.ks; DO klin=1 TO 10; DO klasse = -1.5 TO 1.5 BY 0.1;
    DO bereich = -1 to 1; count = 0; klasse=put(klasse,6.3); OUTPUT; END; END; END;
  RUN;
  DATA SIMUL.klassen; DO klin = 1 TO 10;
    DO klasse = -1.5 TO 1.5 BY 0.1; count = 0; klasse=put(klasse,6.3); OUTPUT; END; END;
  RUN;
  %LET start = 1;
%END;
%IF &neu EQ 0 %THEN %DO;
  DATA SIMUL.marke; SET SIMUL.marke; CALL SYMPUT('letzter',letzter); CALL SYMPUT('h_1',h_1); RUN;
  %LET start = %SYSEVALF(&letzter + 1);
  %LET h_1 = %SYSEVALF(&h_1);
%END;
%LET halten = %sysevalf(&start + &step - 1);
%DO k = &start %TO &start + &lauf - 1;
  PROC DATASETS LIB=WORK; SAVE klin_all blup_all klassen; RUN; QUIT;
  %daten;%auswerten;%organisieren;%ergebnisse; %IF &k EQ &halten %THEN %DO;
    DM 'clear log'; DM 'clear out';
    %gesamtstats; %LET halten = %SYSEVALF(&halten + &step);
  %END;
%END; %mend gesatmakro;

```

## C.3 SAS Programmcode zum Aufbereiten des Datensatzes / Datenersetzungen

```

libname bhir 'C:\...\Daten';

* 1. Schritt. Daten korrigieren und gruppieren;
DATA BHIR.daten_neu; SET DATEN.daten1;
  *Fehler korrigieren;
  IF plz EQ 130347 THEN DO; PLZ = 13347; plz_1 = 2; END;
  IF plz EQ 283955 THEN DO; PLZ = 28395; plz_1 = 3; END;
  IF plz EQ 10 THEN DO; plz_1 = 1; END;
  IF plz EQ 58 THEN DO; plz = 58000; plz_1 = 3; END;
  IF bmi GT 60 THEN DO; bmi = .; END;
  *missings bearbeiten;
  IF alter EQ -99 THEN DO; alter = .; END;
  IF sex EQ -99 THEN DO; sex = .; END;
  IF bmi EQ -99 THEN DO; bmi = .; END;
  IF nation EQ -99 THEN DO; nation = .; END;
  IF beruf EQ -99 THEN DO; beruf = .; END;
  IF plz IN (99, -99, .) THEN DO; PLZ = .; plz_1 = .; END;
  IF famstan2 EQ -99 THEN DO; famstan2 = .; END;
  *neuer Fragebogen;
  ber_rent2= ber_rent;
  IF ber_rent EQ 5 THEN DO; ber_rent2= 6; END;
  IF beruf EQ 5 AND ber_rent2 EQ . THEN DO; berufn = .; END;
  IF beruf EQ 5 AND ber_rent2 EQ -7 THEN DO; ber_rent2= .; END;
  IF beruf NE 5 THEN DO; berufn = beruf; END;
  ELSE DO; berufn = ber_rent2; END;
  IF berufn EQ -99 THEN DO; berufn = .; END;
  *(neu) Gruppieren;
  IF alter LT 30 THEN DO; alterg = 2; END;
  IF alter GE 30 AND alter LT 40 THEN DO; alterg = 3; END;
  IF alter GE 40 AND alter LT 50 THEN DO; alterg = 4; END;
  IF alter GE 50 AND alter LT 60 THEN DO; alterg = 5; END;
  IF alter GE 60 AND alter LT 70 THEN DO; alterg = 6; END;
  IF alter GE 70 AND alter LT 80 THEN DO; alterg = 7; END;
  IF alter GE 80 AND alter LT 90 THEN DO; alterg = 8; END;
  IF alter GE 90 THEN DO; alterg = 9; END;
  IF bmi LT 20 THEN DO; bmi_kl = 1; END;
  IF bmi LT 25 AND bmi GE 20 THEN DO; bmi_kl = 2; END;
  IF bmi LT 30 AND bmi GE 25 THEN DO; bmi_kl = 3; END;
  IF bmi GE 30 THEN DO; bmi_kl = 4; END;
  IF bmi EQ . THEN DO; bmi_kl = .; END;
  IF plz_1 in (1,2) THEN DO; plz_neu = plz_1; END;
  ELSE IF plz IN (
    01941, 01945, 01956, 01957, 01958, 01963, 01964, 01968, 01969, 01971, 01973, 01979, 01980, 01981,
    01983, 01984, 01986, 01987, 01988, 01990, 01991, 01993, 01994, 01996, 01998, 03001, 03002, 03003,
    03004, 03005, 03006, 03007, 03008, 03009, 03010, 03011, 03012, 03013, 03014, 03022, 03024, 03025,
    03029, 03031, 03042, 03044, 03046, 03048, 03050, 03051, 03052, 03053, 03054, 03055, 03058, 03060,
    03062, 03063, 03065, 03094, 03096, 03097, 03099, 03100, 03101, 03103, 03114, 03116, 03117, 03119,
    03121, 03122, 03124, 03130, 03139, 03141, 03142, 03143, 03149, 03157, 03159, 03161, 03162, 03165,
    03172, 03181, 03182, 03185, 03197, 03201, 03205, 03215, 03216, 03217, 03218, 03221, 03222, 03223,
    03226, 03227, 03229, 03231, 03232, 03234, 03235, 03238, 03246, 03248, 03249, 03251, 03252, 03253,
    04891, 04892, 04895, 04906, 04907, 04910, 04911, 04912, 04916, 04920, 04921, 04924, 04928, 04929,
    04931, 04932, 04934, 04936, 04937, 04938, 12529, 12625, 14401, 14402, 14403, 14404, 14405, 14406,
    14407, 14408, 14409, 14410, 14411, 14412, 14413, 14414, 14415, 14416, 14424, 14425, 14427, 14428,
    14429, 14434, 14437, 14438, 14439, 14440, 14443, 14458, 14459, 14460, 14461, 14462, 14463, 14464,
    14467, 14469, 14471, 14473, 14476, 14478, 14480, 14482, 14504, 14505, 14510, 14513, 14525, 14526,
    14532, 14533, 14536, 14542, 14543, 14547, 14548, 14550, 14552, 14554, 14557, 14558, 14601, 14606,
    14609, 14612, 14621, 14624, 14627, 14631, 14632, 14633, 14634, 14641, 14652, 14656, 14658, 14662,
    14664, 14669, 14701, 14702, 14703, 14704, 14705, 14712, 14715, 14723, 14727, 14728, 14731, 14732,
    14733, 14734, 14735, 14736, 14737, 14738, 14745, 14747, 14748, 14749, 14750, 14751, 14765, 14767,

```

```

14770, 14772, 14774, 14776, 14778, 14789, 14790, 14791, 14793, 14794, 14797, 14798, 14801, 14802,
14806, 14819, 14822, 14823, 14824, 14825, 14827, 14828, 14901, 14902, 14903, 14904, 14913, 14926,
14929, 14931, 14932, 14933, 14934, 14937, 14940, 14941, 14943, 14947, 14956, 14959, 14961, 14962,
14963, 14965, 14971, 14974, 14979, 15201, 15202, 15203, 15204, 15205, 15206, 15207, 15208, 15209,
15210, 15225, 15226, 15227, 15228, 15229, 15230, 15232, 15234, 15236, 15292, 15295, 15296, 15299,
15301, 15306, 15316, 15318, 15320, 15322, 15324, 15326, 15327, 15328, 15331, 15332, 15337, 15344,
15345, 15362, 15363, 15364, 15366, 15368, 15370, 15372, 15374, 15377, 15378, 15501, 15502, 15503,
15504, 15505, 15515, 15517, 15518, 15524, 15526, 15528, 15531, 15537, 15558, 15562, 15563, 15566,
15567, 15569, 15701, 15702, 15703, 15706, 15707, 15711, 15728, 15732, 15738, 15739, 15741, 15742,
15743, 15745, 15746, 15748, 15749, 15751, 15752, 15754, 15755, 15757, 15758, 15801, 15806, 15824,
15827, 15828, 15831, 15832, 15834, 15837, 15838, 15841, 15848, 15855, 15859, 15864, 15868, 15871,
15872, 15873, 15874, 15881, 15890, 15896, 15898, 15901, 15902, 15903, 15904, 15905, 15907, 15910,
15913, 15921, 15922, 15926, 15934, 15936, 15938, 16201, 16202, 16203, 16204, 16205, 16206, 16207,
16208, 16212, 16222, 16225, 16227, 16230, 16244, 16247, 16248, 16251, 16252, 16259, 16265, 16269,
16271, 16272, 16278, 16284, 16285, 16286, 16290, 16291, 16292, 16294, 16295, 16303, 16306, 16307,
16311, 16321, 16336, 16341, 16342, 16348, 16349, 16352, 16353, 16356, 16359, 16501, 16502, 16503,
16515, 16535, 16540, 16541, 16542, 16547, 16548, 16552, 16553, 16556, 16558, 16559, 16562, 16565,
16567, 16721, 16727, 16748, 16750, 16752, 16759, 16761, 16763, 16766, 16767, 16771, 16775, 16786,
16792, 16794, 16798, 16801, 16802, 16803, 16812, 16814, 16816, 16818, 16827, 16830, 16831, 16833,
16835, 16837, 16841, 16845, 16856, 16861, 16866, 16868, 16869, 16901, 16905, 16909, 16918, 16921,
16928, 16942, 16945, 16949, 17258, 17261, 17262, 17266, 17268, 17277, 17279, 17281, 17282, 17283,
17291, 17309, 17326, 17335, 17337, 17348, 19307, 19309, 19311, 19312, 19313, 19314, 19322, 19334,
19336, 19337, 19339, 19341, 19342, 19348, 19355, 19357)
THEN DO; plz_neu = 3; END;
ELSE DO; plz_neu = 4; END;
IF plz_1 EQ . THEN DO; plz_neu = .; END;
RUN;

```

## C.4 SAS Programmcode für primäres Analysemodell

```

libname bhir 'C:\...\Daten';

PROC SORT DATA=bhir.daten_neu; BY klinik; RUN;

DATA bhir.modell4 (WHERE = (
  alter NE . AND sex NE . AND pulmstau NE . AND smoker NE . AND hypercho NE . AND arthyp NE .
  AND diabmell NE . AND reani1 NE . )); SET bhir.daten_neu;
RUN;

TITLE1 'Modell 4 - Risikoverteilung';

TITLE2 '1. Alter';
PROC MEANS DATA=bhir.modell4 MAXDEC=2; VAR alter; BY klinik; RUN;
TITLE2 '2. Geschlecht';
PROC FREQ DATA=bhir.modell4; TABLES sex*klinik; RUN;
TITLE2 '3. pulmonaler Stau';
PROC FREQ DATA=bhir.modell4; TABLES pulmstau*klinik; RUN;
TITLE2 '4. Raucher';
PROC FREQ DATA=bhir.modell4; TABLES smoker*klinik; RUN;
TITLE2 '5. HCE';
PROC FREQ DATA=bhir.modell4; TABLES hypercho*klinik; RUN;
TITLE2 '6. AHT';
PROC FREQ DATA=bhir.modell4; TABLES arthyp*klinik; RUN;
TITLE2 '7. DM';
PROC FREQ DATA=bhir.modell4; TABLES diabmell*klinik; RUN;
TITLE2 '8. Reanimationspflicht';
PROC FREQ DATA=bhir.modell4; TABLES reani1*klinik; RUN;

TITLE1 'Auswertung - Hauptanalyse (Modell 4)'; TITLE2;

```

```

PROC GLIMMIX DATA=bhir.modell4;
  CLASS klinik tod sex pulmstau smoker hypercho arthyp diabmell reani1;
  MODEL tod(event='1') = alter sex pulmstau smoker hypercho arthyp diabmell reani1
    / dist=binary link=logit SOLUTION;
  OUTPUT OUT=pred_4 / allstats;
  RANDOM klinik / type=vc SOLUTION CL;
  ODS OUTPUT covparms=WORK.covpar solutionR=WORK.eblup;
RUN;

* Signifikanztest für  $\sigma^2_a$  durchführen;

DATA WORK.covpar; SET WORK.covpar;
  IF estimate NE . AND stderr NE . THEN DO; Z = estimate/stderr; END;
  IF estimate LE 0.0000001 THEN DO; Z = 0; END;
  test_s2a=0; IF Z GT 0.95963281417 THEN DO; test_s2a = 1; END;
RUN;

* Datenexport zur Berechnung erwarteter Letalitätsraten und weiterer Statistiken;

DATA pred_4 (KEEP=klinik tod PredMuPA klin_c tod_c); SET pred_4;
  IF PredMuPA NE . THEN DO; klin_c = klinik; tod_c = tod; END;
RUN;

PROC EXPORT DATA= pred_4 OUTFILE= 'C:\...\out\pred_marg_GLMM.xls';
  DBMS=EXCEL REPLACE; SHEET='model_4';
RUN;

```

## C.5 R Codes zum Aufruf von CART-Analysen

### C.5.1 Analyse demographischer Variablen (Exkurs)

```

library(rpart)
library(foreign)
demog <- read.xport('C:\\...\\Daten\\demog.xpt')
CART1 <- rpart(TOD ~ ALTER + SEX + BMI + NATION + PLZ_NEU + FAMSTAN2 + BERUFN, data=demog)
par(mfrow=c(1,1), mfc col=c(1,1))
plot(CART1, margin=0.1, compress=TRUE)
text(CART1, use.n=TRUE)

```

### C.5.2 Parameterselektion zur Hauptanalyse

```

library(rpart)
library(foreign)
risiko <- read.xport('C:\\...\\Daten\\ranking.xpt')
CART1 <- rpart(TOD ~ ALTER + SEX + PULMSTAU + SMOKER + HYPERCHO + ARTHYP + DIABMELL + REANI1 + SCHOCK
  + FRUEHIN + MANIFINS + NIEREN, data=risiko, na.action=na.exclude)
par(mfrow=c(1,1), mfc col=c(1,1))
plot(CART1, margin=0.1, compress=TRUE)
text(CART1, use.n=TRUE)

```