Qualitätsvergleiche psychiatrischer Einrichtungen

(ANQ – Expertengruppe Methodendiskussion)

Lutz Dümbgen, Anja Mühlemann, Christof Strähl
Universität Bern

Dezember 2018

1 Ausgangslage

Wir betrachten L verschiedende Einrichtungen (Kliniken). Für die k-te Einrichtung sind die Daten zu N_k Behandlungen verfügbar. Für die j-te Behandlung in Einrichtung k stehen uns folgende Informationen zur Verfügung:

$Y_{k,j}$: ein Mass für den Behandlungserfolg
(z.B. HoNOS(ein) − HoNOS(aus)),

$X_{k,j}$: eine Liste von Kovariablen, welche Patient(in) und Umstände der Behandlung beschreiben
(z.B. Geschlecht, Alter, HoNOS(ein), Diagnose, ...).

Das Ziel ist nun, die Qualität der Einrichtungen in Bezug auf die Werte $Y_{k,j}$ zu vergleichen.

2 Das Modell

Essentially, all models are wrong, but some are useful. (G.E.P. Box)

Würde man die Zusatzinformationen $X_{k,j}$ einfach ignorieren, könnte man die Mittelwerte

$$\bar{Y}_k \;=\; \frac{1}{N_k}\sum_{j=1}^{N_k} Y_{k,j}$$

berechnen und miteinander vergleichen. Es besteht jedoch Konsens darüber, dass so Äpfel mit Birnen verglichen würden, da die in $X_{k,j}$ erfassten Eigenschaften potentiell erheblichen Einfluss auf den Behandlungserfolg $Y_{k,j}$ haben. Die Idee ist, dass

$$\begin{aligned} Y_{k,j} &= f(X_{k,j}) + Y_{k,j}^{\mathrm{adj}} \\ &= f(X_{k,j}) + \mu_k + \epsilon_{k,j} \end{aligned} \tag{1}$$

mit einer gewissen Funktion f der Zusatzinformationen, die nicht von den speziellen Einrichtungen abhängt, gewissen Qualitätsparametern $\mu_1, \mu_2, \ldots, \mu_L$ der Einrichtungen und zufälligen Schwankungen ("Rauschen") $\epsilon_{k,j}$, die aber im Mittel gleich Null sind. Man zerlegt den Behandlungserfolg $Y_{k,j}$ also in drei Teile:

(i) Einen systematischen Anteil $f(X_{k,j})$, der von messbaren und erfassten Eigenschaften der Patient(inn)en und speziellen Umständen der Behandlung abhängt, welche von der Einrichtung nicht beeinflusst werden können.

(ii) Einen systematischen Anteil μ_k, der ausschließlich von der Einrichtung abhängt.

(iii) Einen zufälligen Anteil, der Unterschiede von Patient(in) zu Patient(in) beschreibt, welche nicht in $X_{k,j}$ erfasst sind, zufällige Schwankungen in der persönlichen Verfassung oder Leistung der an der Behandlung beteiligten Personen, und vieles mehr. Die Hoffnung ist, dass $X_{k,j}$ alle für die Qualitätsbeurteilung relevanten Kovariablen enthält; ansonsten enthielte $\epsilon_{k,j}$ auch systematische Anteile.

Damit diese Zerlegung von $Y_{k,j}$ eindeutig ist, sollte man beispielsweise fordern, dass

$$\sum_{k=1}^{L} \sum_{j=1}^{N_k} f(X_{k,j}) \;=\; 0.$$

Dann hat μ_k folgende Interpretation: Es ist der mittlere Behandlungserfolg, wenn alle $N = N_1 + N_2 + \cdots + N_L$ Behandlungen in Einrichtung k durchgeführt würden.

Betreffend f bietet sich im einfachsten Fall ein *multiples lineares Modell* an: Angenommen, die Zusatzinformationen wurden so kodiert, dass $X_{k,j}$ aus numerischen oder $\{0,1\}$-wertigen Grössen $X_{k,j}^{(1)}, X_{k,j}^{(2)}, \ldots, X_{k,j}^{(p)}$ besteht. Dann geht man davon aus, dass

$$f(X_{k,j}) \;=\; \sum_{\ell=1}^{p} \beta_\ell \widetilde{X}_{k,j}^{(\ell)} \tag{2}$$

mit gewissen Parametern $\beta_1, \beta_2, \ldots, \beta_p$ und den zentrierten Kovariablen

$$\widetilde{X}_{k,j}^{(\ell)} \;=\; X_{k,j}^{(\ell)} - \bar{X}^{(\ell)}, \quad \bar{X}^{(\ell)} \;=\; \frac{1}{N} \sum_{k=1}^{L} \sum_{j=1}^{N_k} X_{k,j}^{(\ell)}.$$

Jeder einzelne Koeffizient β_ℓ hat folgende Bedeutung: Ändert man bei einer Behandlung die Kovariable $X_{k,j}^{(\ell)}$ um $\pm$ eine Einheit, während alle anderen Kovariablen unverändert bleiben, dann nimmt der Behandlungserfolg im Mittel um $\pm\beta_\ell$ zu bzw. ab.

Alternativ könnte man auch ein *multiples lineares Modell mit Interaktionen* annehmen:

$$f(X_{k,j}) \;=\; \sum_{\ell=1}^{p} \beta_\ell \widetilde{X}_{k,j}^{(\ell)} \;+\; \sum_{(\ell,m)} \beta_{\ell,m} \widetilde{X}_{k,j}^{(\ell,m)}, \tag{3}$$

wobei sich letztere Summe über alle oder zumindest einige Indexpaare (ℓ, m) mit $1 \le \ell < m \le p$ erstreckt, und

$$\widetilde{X}_{k,j}^{(\ell,m)} \;=\; X_{k,j}^{(\ell)} X_{k,j}^{(m)} - c_{\ell,m} X_{k,j}^{(\ell)} - d_{\ell,m} X_{k,j}^{(m)} - e_{\ell,m}$$

mit gewissen Koeffizienten $c_{\ell,m}, d_{\ell,m}, e_{\ell,m}$ derart, dass $\sum_{k,j} \widetilde{X}_{k,j}^{(\ell,m)} = \sum_{k,j} \widetilde{X}_{k,j}^{(\ell,m)} X_{k,j}^{(\ell)} = \sum_{k,j} \widetilde{X}_{k,j}^{(\ell,m)} X_{k,j}^{(m)} = 0$.

Einen solchen Beitrag $\beta_{\ell,m} \widetilde{X}_{k,j}^{(\ell,m)}$ nennt man Interaktion zwischen der ℓ-ten und der m-ten Kovariable. Interaktionen bieten sich an, wenn man davon ausgeht, dass der Einfluss einer bestimmten Kovariable auf den Behandlungserfolg vom Wert anderer Kovariablen abhängt.

3 Auswertung der Daten

Die nachfolgend beschriebenen Methoden basieren überwiegend auf Standardwerkzeugen wie sie z.B. in [3] beschrieben werden. Weiterführende Themen wie beispielsweise simultane Konfidenzbereiche oder "Wild bootstrap" werden in den Vorlesungsskripten [1] ausführlich behandelt. Das Hauptziel sind statistisch zuverlässige Aussagen über die Qualitätsparameter μ_k sowie die daraus abgeleiteten Vergleichsgrössen

$$\delta_k \;=\; \mu_k - \frac{1}{N - N_k} \sum_{k' \neq k} N_{k'} \mu_{k'},$$

also die Differenzen zwischen jedem einzelnen μ_k und dem mit den Fallzahlen gewichteten Mittel der übrigen Parameter $\mu_{k'}$, $k' \neq k$. Dass δ_k grösser oder kleiner als 0 ist, deuten wir dahingehend, dass die Behandlungserfolge in Einrichtung k überdurchschnittlich hoch bzw. niedrig sind.

Kleinste-Quadrate-Schätzung. Ausgehend von der Modellgleichung (1) und der konkreten Form (2) oder (3) von $f(X_{k,j})$ kann man sowohl die Funktion f als auch die Qualitätsparameter μ_k simultan schätzen: Man wählt Parameter $\widehat{\mu}_1, \widehat{\mu}_2, \ldots, \widehat{\mu}_L$ sowie eine Funktion $\widehat{f}$ vom Typ (2) bzw. (3), so dass die Quadratsumme

$$\sum_{k=1}^{L} \sum_{j=1}^{N_k} (Y_{k,j} - \widehat{\mu}_k - \widehat{f}(X_{k,j}))^2$$

möglichst klein ist.

Unter der Annahme, dass die zufälligen Schwankungen $\epsilon_{k,j}$ wirklich Erwartungswert 0 haben, stellen die so geschätzten Qualitätsmasse $\widehat{\mu}_1, \widehat{\mu}_2, \ldots, \widehat{\mu}_L$ unverzerrte Schätzer für die tatsächlichen Grössen $\mu_1, \mu_2, \ldots, \mu_L$ dar. Das heisst, sie enthalten keine systematischen Fehler.

Standardfehler und Konfidenzintervalle. Unter der zusätzlichen Annahme, dass die Zufallsgrössen $\epsilon_{k,j}$ unabhängig sind und einheitliche Standardabweichung σ haben, kann man die unvermeidliche Ungenauigkeit der Grössen $\widehat{\mu}_k$ oder $\widehat{\delta}_k$ durch entsprechende

Standardfehler (ihre geschätzten Standardabweichungen) quantifizieren. Ferner lassen sich für die entsprechenden δ_k einfache und simultane Vertrauensintervalle angeben. Letztlich führt dies zu einer Tabelle der Form

k	$\widehat{\mu}_k$	$\mathrm{SE}(\widehat{\mu}_k)$	$\widehat{\delta}_k$	$\mathrm{SE}(\widehat{\delta}_k)$	a_k	b_k	a_k^{sim}	b_k^{sim}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

bzw. graphischen Darstellungen dieser Einträge. Dabei sind $\mathrm{SE}(\widehat{\mu}_k)$ und $\mathrm{SE}(\widehat{\delta}_k)$ die Standardfehler von $\widehat{\mu}_k$ bzw. $\widehat{\delta}_k$, und $[a_k, b_k]$, $[a_k^{\mathrm{sim}}, b_k^{\mathrm{sim}}]$ sind einfache bzw. simultane Konfidenzschranken für die Parameter δ_k: Für ein vorgegebenes Vertrauensniveau $1-\alpha$ (in der Regel 95%) ist

$$P(a_k \leq \delta_k \leq b_k) \;=\; 1-\alpha$$

für jede einzelne Einrichtung k, und

$$P\big(a_k^{\mathrm{sim}} \leq \delta_k \leq b_k^{\mathrm{sim}} \;\text{ für } k = 1, 2, \ldots, L\big) \;=\; 1-\alpha.$$

Die einzelnen Konfidenzintervalle $[a_k, b_k]$ werden mit der üblichen Student-Methode berechnet, die simultanen Intervalle $[a_k^{\mathrm{sim}}, b_k^{\mathrm{sim}}]$ mit einer Verallgemeinerung der Tukey-Methode (verwandt mit studentized ranges in ANOVA).

Residuenanalyse. Die zuletzt genannten Vertrauensintervalle basieren auf der zusätzlichen Annahme, dass die $\epsilon_{k,j}$ normalverteilt sind oder die "Hebelwirkungen" des Regressionsmodells klein sind. All diese Annahmen kann und sollte man mit entsprechenden graphischen und numerischen Methoden überprüfen.

Wild bootstrap. Falls die Residuenanalysen zeigen, dass die Standardabweichungen der $\epsilon_{k,j}$ deutlich von $f(X_{k,j})$ abhängen, bietet sich als Alternative bzw. zur Ergänzung das "Wild Bootstrap" für lineare Regression an. Dabei handelt es sich um eine etablierte Resampling-Methode, die unter deutlich schwächeren Bedingungen zumindest approximativ valide Standardfehler und Vertrauensintervalle liefert.

Auswahl von Kovariablen. In der vorliegenden Anwendung ist mit enorm grossen Fallzahlen im Bereich von mehreren Tausend zu rechnen. Von daher empfehlen wir folgendes Vorgehen: Die Experten aus dem klinischen Bereich einigen sich auf eine Kollektion von potentiell relevanten numerischen und kategoriellen Kovariablen, die zu berücksichtigen sind. Wünschenswert wäre eine Liste von bis zu ca. zehn Merkmalen.

There is no substitute for good data, and there is no substitute for expertise!
(W.F. Eddy)

Nach Kodierung der kategoriellen Kovariablen durch $\{0,1\}$-wertige Indikatoren ergeben sich letztlich p Kovariablen und $\binom{p}{2} = p(p-1)/2$ Interaktionen, plus die L Klinikparameter,

also $d = L + p(p+1)/2$ Parameter. Die Fallzahl n sollte mindestens $5(L+p)$ betragen, damit man zumindest das einfache Modell (2) gut anwenden kann.

Falls die Anzahl n von Fällen mindestens $5d$ ist, wird das in (3) beschriebene multiple lineare Modell mit allen Interaktionen verwendet. Sollte sich herausstellen, dass die entsprechende Design-Matrix schlecht koniditioniert ist, das heisst, sollten gewisse Spalten (fast) linear abhängig von den übrigen sein, werden solche schrittweise entfernt und das Modell dementsprechend reduziert.

Falls es zwar mindestens $5(L+p)$ aber weniger als $5d$ Fälle gibt, startet man mit dem einfacheren multiplen linearen Modell (2). Danach werden durch Vorwärtsselektion solche Interaktionen hinzugefügt, welche signifikanten Einfluss zeigen. Diese Situation könnte beispielsweise bei den Fällen aus Kinder- und Jugendpsychiatrie auftreten.

Nebenbemerkung: Als Ergänzung zur Vorwärtsselektion von Interaktionen, aber auch zum automatischen Auswählen von relevanten Kovariablen aus einer grösseren Kollektion bietet sich das LASSO-Verfahren (least absolute sum selection operator, [4]) an: Dieses liefert bei im Prinzip beliebig grosser Anzahl von Kovariablen und Interaktionen eine Rangfolge der potentiell relevanten Kovariablen und Interaktionen. In einem zweiten Schritt könnte man dann mittels forward selection die letztlich zu verwendenden Kovariablen und Interaktionen auswählen.

Genauer gesagt, bestimmt man für einen Regularisierungsparameter $\lambda > 0$ eine Regressionsfunktion $\widehat{f}_\lambda$ vom Typ (2) (oder (3)) und Koeffizienten $\widehat{\mu}_{1,\lambda}, \widehat{\mu}_{1,\lambda}, \ldots, \widehat{\mu}_{L,\lambda}$, so dass die Summe

$$\sum_{j,k}(Y_{k,j} - \widehat{\mu}_{k,\lambda} - \widehat{f}_\lambda(X_{k,j}))^2 + \lambda \sum_\ell s_\ell |\widehat{\beta}_\ell| \quad \Big(+ \lambda \sum_{(\ell,m)} s_\ell s_m |\widehat{\beta}_{\ell,m}|\Big)$$

möglichst klein wird. Dabei ist

$$s_\ell \;=\; \sqrt{\sum_{j,k}(X_{k,j}^{(\ell)} - \bar{X}^{(\ell)})^2}.$$

Man ergänzt also die übliche Quadratsumme durch einen Bestrafungsterm mal λ. Dieser verursacht, dass die Koeffizienten $\beta_\ell, \beta_{\ell,m}$ tendenziell zu 0 hin verschoben werden und manche sogar gleich 0 sind. Mit wachsendem λ werden mehr und mehr Koeffizienten gleich 0 gesetzt. Die potentielle Relevanz einer Kovariable oder Interaktion kann man durch den kleinsten Wert λ quantifizieren, für welchen ihr Parameter β_ℓ bzw. $\beta_{\ell,m}$ noch gleich 0 ist.

Rückmeldungen an einzelne Kliniken. Jede einzelne Klinik k sollte nicht nur mit den Schätzwerten $\widehat{\mu}_k$, $\widehat{\delta}_k$ und einem Vertrauensintervall für δ_k "abgespeist" werden. Um der Klinik das Aufspüren von aussergewöhnlichen Fällen zu erleichtern, könnte man ihr

eine ergänzte Datenmatrix der folgenden Form zurückgeben:

Fallnr.	Kovariablen		Erfolgsmass	erw. Erfolgsm.	Abweichung
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
j	$X_{kj}^{(1)}$ $\dots$	$X_{kj}^{(p)}$	Y_{kj}	$\widehat{Y}_{kj}^{o}$	$Y_{kj} - \widehat{Y}_{kj}^{o}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

Dabei würde j alle N_k Fälle von Klinik k durchlaufen, und

$$\widehat{Y}_{kj}^{o} \;=\; \sum_{k'=1}^{L} \frac{N_{k'}}{N}\, \widehat{\mu}_k + \widehat{f}(X_{kj})$$

wäre der geschätzte erwartete Behandlungserfolg eines Falles vom Typ X_{kj}, gemittelt über alle Kliniken. Die N_k Fälle könnte man nach der Grösse der Abweichung $Y_{kj} - \widehat{Y}_{kj}^{o}$ anordnen.

Eine solche Tabelle würde der Klinik erleichtern herauszufinden, welche Fälle sie nach Augenschein besonders gut oder eher schlecht behandelte.

Zusätzlich könnte man die Abweichungen $Y_{kj} - \widehat{Y}_{kj}^{o}$ gegen gewisse numerische Kovariablen auftragen oder Box-Plots der Abweichungen als Funktion gewisser kategorieller Kovariablen erzeugen. Auf diese Weise könnte die Klinik herausfinden, wo ihre Stärken und Schwächen liegen.

Denkbar ist auch, dass mit Hilfe solcher Listen und Graphiken relevante Kovariablen, die bisher nicht berücksichtigt wurden, erkannt werden. Somit würde auch eine Weiterentwicklung des Regressionsmodells unter Beteiligung der Kliniken unterstützt.

4 Illustration

Wir illustrieren die Auswertungsmethode bzw. deren Resultate an einem Testdatensatz, welcher ca. 10% der tatsächlich vorhandenen Daten umfasst. Aufgrund der teilweise geringen Fallzahlen unterscheiden wir keine Kliniktypen. Wir verwendeten das lineare Modell (3) mit der Response $Y = \text{HoNOS(ein)} - \text{HoNOS(aus)}$ und folgenden Kovariablen: Geschlecht, Alter, HoNOS(ein) und Hauptdiagnose. Letztere wurde als kategorielles Merkmal mit neun Ausprägungen, also acht $\{0,1\}$-wertigen Kovariablen kodiert. Bereits bei diesem Teildatensatz zeigte sich, dass das Modell 3 mit Interaktionen signifikant besser ist als das einfachere lineare Modell 2; ein entsprechender F-Test lieferte den P-Wert 0.0264.

Die Tabellen 1 und 2 enthalten die Resultate dieser Auswertung (alle Werte auf drei Nachkommastellen gerundet) mit Vertrauensniveau 95%. Bei 5 Einrichtungen ist $a_k^{\text{sim}} > 0$, und bei 5 anderen Einrichtungen ist $b_k^{\text{sim}} < 0$. Bei den übrigen 52 Einrichtungen war $a_k^{\text{sim}} < 0 < b_k^{\text{sim}}$. Mit einer Sicherheit von 95% kann man behaupten, dass $\delta_k > 0$ für alle Einrichtungen mit $a_k^{\text{sim}} > 0$ und dass $\delta_k < 0$ für alle Einrichtungen mit $b_k^{\text{sim}} < 0$. Der

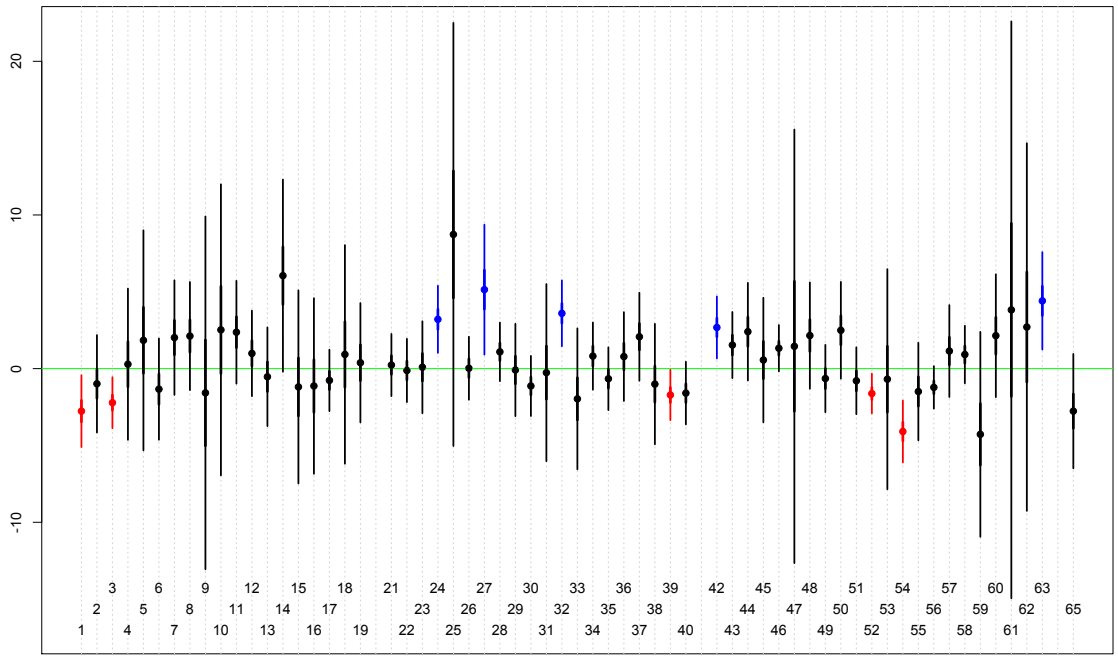

Abbildung 1: Auswertungsbeispiel: Schätzer und Konfidenzbereiche für δ_k.

Schätzwert $\widehat{\bar{\mu}}$ für $\bar{\mu}$ ist gleich 6.483. Letzteres ist der erwartete durchschnittliche Behandlungserfolg, gemittelt über alle Falltypen und Kliniken.

Eine graphische Darstellung dieser Ergebnisse sieht man in Abbildung 1. Für alle 62 Einrichtungen mit verfügbaren Daten sieht man den Punktschätzer $\widehat{\delta}_k$ und das Intervall $[a_k^{\text{sim}}, b_k^{\text{sim}}]$. Ausserdem wird noch der Bereich $\left[\widehat{\delta}_k \pm \text{SE}(\widehat{\delta}_k)\right]$ etwas hervorgehoben. Diejenigen Intervalle $[a_k^{\text{sim}}, b_k^{\text{sim}}]$, welche den Wert 0 nicht enthalten, sind rot ($b_k^{\text{sim}} < 0$) bzw. blau ($a_k^{\text{sim}} > 0$) gezeichnet.

Zuguterletzt illustrieren wir noch, wie man die Ergebnisse klinikweise analysieren kann. Und zwar zeigen wir in den Abbildungen 3 für die fünf potentiell problematischen und in Abbildung 4 für die fünf potentiell herausragenden Kliniken jeweils Box-Whisker-Plots der Differenzen $Y_{k,j} - \widehat{Y}^o_{k,j}$, nach Hauptdiagnosen getrennt. Zum Vergleich werden auch die Box-Whisker-Plots dieser Differenzen in allen 62 Kliniken gezeichnet. Auffallend ist, dass bei vielen der hier betrachteten Kliniken die Qualität durchaus von der Diagnosegruppe abhängt. Bei der Suche nach möglichen Ursachen könnten solche Analysen hilfreich sein.

5 Schlussbemerkung

Bei der Präsentation der Resultate mit der hier vorgeschlagenen Methode sollte man im Auge behalten und erwähnen, dass es einige Unsicherheiten gibt, die man auch mit noch so schönen statistischen Methoden nicht ausschliessen kann. Die geschätzten Qualitätspa-

k	$\widehat{\mu}_k$	$\mathrm{SE}(\widehat{\mu}_k)$	$\widehat{\delta}_k$	$\mathrm{SE}(\widehat{\delta}_k)$	a_k	b_k	a_k^{sim}	b_k^{sim}
1	3.773	0.695	−2.767	0.702	−4.144	−1.390	−5.108	**−0.426**
2	5.507	0.947	−0.987	0.952	−2.854	0.880	−4.160	2.186
3	4.357	0.488	−2.216	0.498	−3.193	−1.238	−3.877	**−0.555**
4	6.768	1.474	0.286	1.478	−2.611	3.183	−4.638	5.210
5	8.319	2.147	1.840	2.149	−2.373	6.054	−5.322	9.003
6	5.161	0.984	−1.336	0.989	−3.275	0.604	−4.633	1.961
7	8.486	1.114	2.019	1.119	−0.175	4.213	−1.710	5.748
8	8.585	1.052	2.120	1.056	0.049	4.191	−1.400	5.640
9	4.907	3.444	−1.577	3.445	−8.333	5.178	−13.060	9.905
10	9.001	2.841	2.522	2.842	−3.051	8.094	−6.951	11.994
11	8.827	0.998	2.369	1.004	0.401	4.337	−0.976	5.714
12	7.459	0.830	0.990	0.836	−0.649	2.629	−1.796	3.776
13	5.959	0.959	−0.529	0.964	−2.420	1.362	−3.743	2.685
14	12.510	1.872	6.049	1.877	2.369	9.728	−0.205	12.303
15	5.295	1.883	−1.191	1.886	−4.888	2.507	−7.476	5.095
16	5.357	1.713	−1.130	1.715	−4.493	2.234	−6.847	4.587
17	5.737	0.592	−0.767	0.601	−1.945	0.411	−2.769	1.235
18	7.405	2.133	0.924	2.136	−3.263	5.111	−6.193	8.041
19	6.864	1.162	0.384	1.166	−1.902	2.670	−3.502	4.269
21	6.710	0.599	0.234	0.608	−0.958	1.425	−1.791	2.258
22	6.366	0.611	−0.120	0.619	−1.333	1.093	−2.182	1.941
23	6.573	0.893	0.091	0.898	−1.670	1.852	−2.903	3.085
24	9.618	0.649	3.208	0.656	1.921	4.495	**1.020**	5.396
25	15.209	4.130	8.732	4.131	0.632	16.832	−5.036	22.500
26	6.507	0.608	0.024	0.616	−1.183	1.232	−2.028	2.077
27	11.593	1.265	5.141	1.269	2.652	7.629	**0.911**	9.371
28	7.540	0.563	1.091	0.572	−0.031	2.213	−0.817	2.998
29	6.394	0.896	−0.090	0.902	−1.857	1.678	−3.094	2.915
30	5.387	0.578	−1.129	0.587	−2.281	0.023	−3.087	0.829
31	6.223	1.727	−0.260	1.730	−3.652	3.132	−6.026	5.505
32	9.994	0.635	3.598	0.643	2.338	4.859	**1.456**	5.740

Tabelle 1: Auswertungsbeispiel, Teil I

k	$\widehat{\mu}_k$	$\mathrm{SE}(\widehat{\mu}_k)$	$\widehat{\delta}_k$	$\mathrm{SE}(\widehat{\delta}_k)$	a_k	b_k	a_k^{sim}	b_k^{sim}
33	4.525	1.373	−1.968	1.376	−4.666	0.730	−6.554	2.618
34	7.279	0.649	0.815	0.657	−0.473	2.103	−1.374	3.004
35	5.842	0.605	−0.658	0.613	−1.860	0.544	−2.701	1.385
36	7.257	0.863	0.784	0.868	−0.919	2.487	−2.110	3.678
37	8.529	0.855	2.073	0.860	0.387	3.760	−0.794	4.940
38	5.483	1.172	−1.008	1.177	−3.314	1.299	−4.929	2.913
39	4.843	0.481	−1.715	0.492	−2.680	−0.749	−3.355	**−0.074**
40	4.933	0.604	−1.594	0.612	−2.795	−0.393	−3.635	0.448
42	9.089	0.598	2.680	0.607	1.490	3.869	**0.658**	4.702
43	7.981	0.639	1.535	0.646	0.267	2.802	−0.620	3.689
44	8.860	0.949	2.403	0.954	0.532	4.275	−0.777	5.584
45	7.039	1.212	0.559	1.216	−1.826	2.945	−3.495	4.614
46	7.743	0.441	1.328	0.453	0.440	2.216	−0.182	2.837
47	7.933	4.232	1.451	4.233	−6.850	9.751	−12.658	15.559
48	8.616	1.034	2.152	1.038	0.117	4.188	−1.308	5.613
49	5.852	0.650	−0.646	0.658	−1.936	0.645	−2.838	1.547
50	8.949	0.940	2.494	0.946	0.640	4.348	−0.657	5.646
51	5.710	0.646	−0.791	0.654	−2.073	0.490	−2.969	1.387
52	4.975	0.375	−1.621	0.389	−2.383	−0.859	−2.917	**−0.325**
53	5.797	2.147	−0.688	2.150	−4.903	3.527	−7.852	6.476
54	2.503	0.597	−4.092	0.605	−5.279	−2.905	−6.109	**−2.075**
55	5.009	0.948	−1.491	0.954	−3.361	0.378	−4.669	1.687
56	5.336	0.402	−1.220	0.415	−2.033	−0.406	−2.603	0.163
57	7.613	0.892	1.144	0.898	−0.616	2.904	−1.847	4.135
58	7.373	0.553	0.919	0.562	−0.184	2.021	−0.956	2.793
59	2.216	1.999	−4.277	2.001	−8.201	−0.353	−10.947	2.393
60	8.611	1.196	2.142	1.200	−0.211	4.496	−1.858	6.143
61	10.303	5.632	3.821	5.633	−7.223	14.866	−14.951	22.594
62	9.186	3.589	2.706	3.591	−4.335	9.746	−9.262	14.673
63	10.845	0.946	4.410	0.951	2.545	6.275	**1.240**	7.580
65	3.739	1.113	−2.766	1.118	−4.958	−0.574	−6.491	0.959

Tabelle 2: Auswertungsbeispiel, Teil II

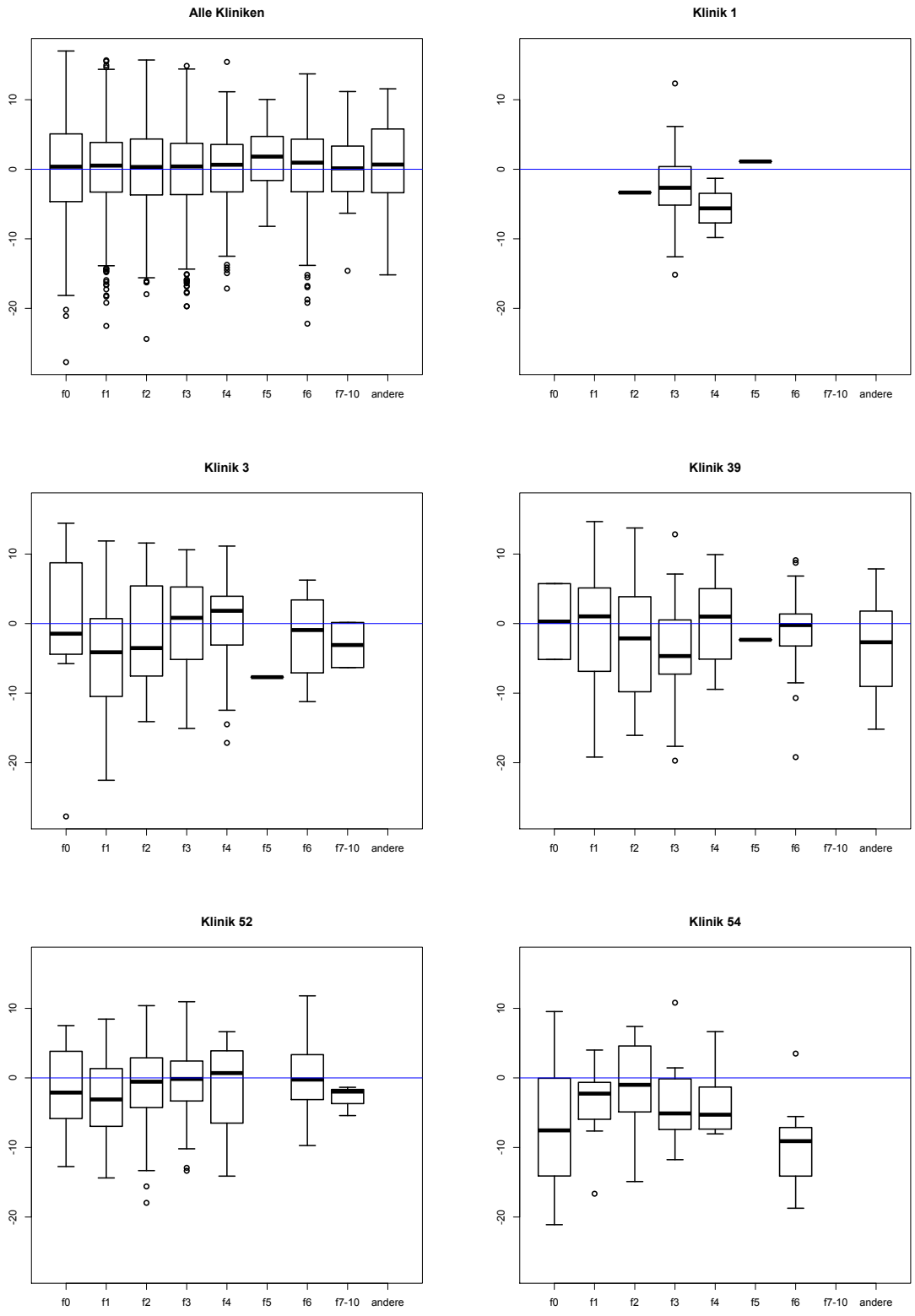

Abbildung 2: Auswertungsbeispiel: Abweichungen der Behandlungserfolge von den erwarteten Werten in potenziell problematischen Kliniken.

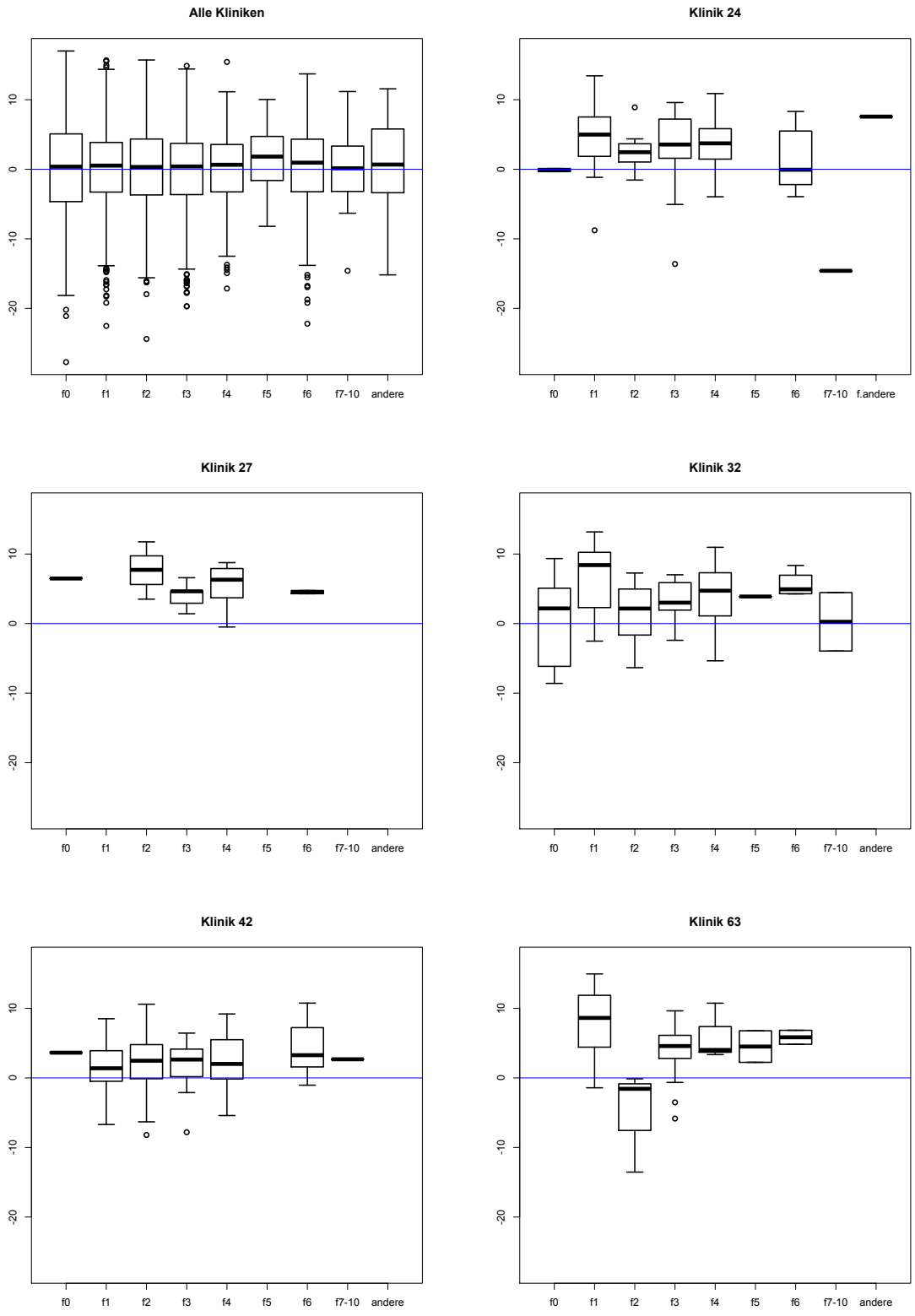

Abbildung 3: Auswertungsbeispiel: Abweichungen der Behandlungserfolge von den erwarteten Werten in potenziell herausragenden Kliniken.

rameter $\widehat{\mu}_k$ und $\widehat{\delta}_k$ können aus zweierlei Gründen systematische Fehler enthalten:

(i) Wichtige Kovariablen wurden nicht berücksichtigt, da die entsprechenden Daten nicht in ausreichender Qualität verfügbar waren, oder weil noch niemand an sie dachte.

(ii) Der Einfluss der Kovariablen wird durch das Modell (2) oder (3) nicht adäquat beschrieben.

Punkt (ii) wird zwar in seriös durchgeführten Residuenanalysen überprüft, aber bei Modellen mit vielen Kovariablen können solche Abweichungen unerkannt bleiben.

Ferner sollte man keinesfalls die geschätzten Parameter $\widehat{\mu}_k$ oder $\widehat{\delta}_k$ ohne Angabe von Standardfehlern oder Vertrauensschranken publizieren und beispielsweise für ein einfaches Ranking heranziehen. Die hier vorgeschlagene Methode erscheint uns vor allem als Instrument für ein Screening der Einrichtungen geeignet. Man identifiziert Einrichungen, welche augenscheinlich besonders gut oder suboptimal arbeiten und die man allenfalls genauer unter die Lupe nehmen sollte.

Literatur

[1] L. Dümbgen (2015). *Lineare Modelle und Regression.* Lecture notes, University of Bern.

[2] Lutz Dümbgen (2016). (Ab)Using Regression Methods for Data Adjustment. Technical report 78, IMSV, University of Bern.

[3] T.P. Ryan (1997). *Modern Regression Methods.* Wiley, New York.

[4] R. Tibshirani (1996). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society, Series B* ***58***, pp. 267-288.