IQTIG

Institut für Qualitätssicherung und Transparenz im Gesundheitswesen

Volumen-Outcome-Beziehungen bei Revisionseingriffen in der Knieendoprothetik

Sonderauswertung des vormaligen Leistungsbereichs Knie-Endoprothesenwechsel und -komponentenwechsel

Erstellt im Auftrag des
Gemeinsamen Bundesausschusses

Stand: 20. Dezember 2019

Impressum

Thema:
Volumen-Outcome-Beziehungen bei Revisionseingriffen in der Knieendoprothetik. Sonderauswertung des vormaligen Leistungsbereichs Knie-Endoprothesenwechsel und -komponentenwechsel

Ansprechpartner:
Dr. Arne Deiseroth

Auftraggeber:
Gemeinsamer Bundesausschuss

Datum des Auftrags:
16. Mai 2019

Datum der Abgabe:
20. Dezember 2019

Herausgeber:
IQTIG – Institut für Qualitätssicherung
und Transparenz im Gesundheitswesen

Katharina-Heinroth-Ufer 1
10787 Berlin

Telefon: (030) 58 58 26-0
Telefax: (030) 58 58 26-999

info@iqtig.org
https://www.iqtig.org

Inhaltsverzeichnis

Tabellenverzeichnis

Abbildungsverzeichnis

Glossar mathematischer Notationen

Notation	Erläuterung
Zähl-Indizes	
$i = 1 \dots, I$	Index für die Leistungserbringer
$j = 1, \dots, J_i$	Index für die Patientinnen und Patienten von Leistungserbringer i
Variablen des Inferenzmodells (vgl. Gleichung (1) bzw. Abbildung 5)	
y_{ij}	Binäres Outcome für Patientin oder Patient j behandelt von Leistungserbringer i bzgl. eines Qualitätsindikators
η_{ij}	Logit-transformierte Wahrscheinlichkeit für das Eintreten des Ereignisses $y_{ij} = 1$, geschätzt auf Basis des jeweiligen Risikoadjustierungsmodells
n_i	Fallzahl: Anzahl an Knieendoprothesen-Wechselprozeduren von Leistungserbringer i im Erfassungsjahr 2014. Abweichend für QI 51874 und QI 51044 die Anzahl an Fällen in der Grundgesamtheit des jeweiligen Indikators.
$f(n_i)$	Kontinuierliche Funktion der Fallzahl zur Modellierung des Fallzahl-Ergebnis-Zusammenhangs
u_i	Random-Intercept zur Modellierung der fallzahlunabhängigen Grundkompetenz von Leistungserbringer i
τ	Standardabweichung der als a-priori normalverteilt angenommenen Random-Intercepts u_i
β_0	Globaler Intercept-Parameter
Knoten des Kausalmodelles (Abbildung 3)	
$\boldsymbol{x}_{ij} = (x_{ij}^1, \dots, x_{ij}^M)$	Beobachtete potentielle patientenseitige Risikofaktoren für ein betrachtetes Patienten-Outcome
c_{ij}	Unbeobachtete (sonstige) potentielle Risikofaktoren für ein betrachtetes Patienten-Outcome
b_i	Behandlungskompetenz von Leistungserbringer i im betrachteten Qualitätsaspekt
$\boldsymbol{z_i} = \left(z_i^1, \dots, z_i^L\right)$	Beobachtete Leistungserbringermerkmale von Leistungserbrigner i mit Kausaleinfluss auf die Behandlungskompetenz b_i im betrachteten Qualitätsaspekt

Notation	Erläuterung
s_i	Unbeobachtete (sonstige) Leistungserbringermerkmale von Leistungserbringer i mit Kausaleinfluss auf die Behandlungskompetenz b_i im betrachteten Qualitätsaspekt
Allgemeine mathematische Notation	
$\mathrm{Log}(x)$	Natürlicher Logarithmus von x
$\mathrm{Logit}(x)$	Logit-Transformation von x: $\mathrm{Logit}(x) = \mathrm{Log}\left(\frac{x}{1-x}\right)$ für $x \in (0,1)$
$\mathrm{Logit}^{-1}(y)$	Inverse Logit-Transformation: $\mathrm{Logit}^{-1}(y) = \frac{\exp(y)}{1+\exp(y)}$ für $y \in (-\infty, \infty)$
$\mathrm{Odds}(y = 1\|z)$	Odds: Verhältnis von Wahrscheinlichkeit und Gegenwahrscheinlichkeit für das Ereignis $y = 1$ gegeben einen Wert z: $\frac{\mathrm{P}(y = 1\|z)}{1-\mathrm{P}(y = 1\|z)}$
$\mathrm{MOR}(\tau)$	Median-Odds-Ratio: (vgl. Gleichung (2) in Abschnitt 4.1.3) $\mathrm{MOR}(\tau) = \exp(\sqrt{2 \cdot \tau^2} \cdot \Phi^{-1}(3/4))$
$h \equiv g$	Gleichheit der Funktionen h und g an allen Punkten des Definitionsbereiches.
$h \not\equiv g$	Die Funktionen h und g sind nicht an allen Punkten des Definitionsbereiches gleich.
$\int_a^b f(x)\,\mathrm{d}x$	Integral der Funktion f über dem Intervall [a,b].
$f''(t)$	zweite Ableitung der Funktion f
$\hat{x}$	Datenbasierte Schätzung von Variable x
$X \perp Y$	Zufallsvariablen X und Y sind statistisch unabhängig
$X \mid Y$	Zufallsvariable X bedingt auf Zufallsvariable Y
$X \sim F$	Zufallsvariable X folgt der Wahrscheinlichkeitsverteilung F
$X_i \overset{\text{u.i.v.}}{\sim} F, i = 1, \ldots, n$	Zufallsvariablen $X_1, \ldots, X_n$ sind unabhängig und identisch verteilt und folgen der Wahrscheinlichkeitsverteilung F
$\mathrm{Ber}(\pi)$	Bernoulli-Verteilung mit Erfolgswahrscheinlichkeit $\pi \in [0,1]$
$\mathcal{N}(\mu, \sigma^2)$	Normalverteilung mit Erwartungswert μ und Varianz σ^2
$\Phi(z)$	Verteilungsfunktion der Standard-Normalverteilung

Zusammenfassung

Hintergrund

Die „Volumen-Outcome-Hypothese“ geht zurück auf Luft et al. (1979) und beruht auf der Annahme eines Zusammenhangs zwischen der Menge erbrachter Leistungen eines Krankenhauses (z. B. Operationen) und dem Behandlungsergebnis (z. B. Sterblichkeit). Zahlreiche Primär- und Registerstudien haben in den letzten Jahrzenten die Beziehung zwischen Volumen und Outcome in der Knieendoprothetik untersucht. Insbesondere für fachlich anspruchsvolle Eingriffe wie den Wechsel bereits liegender Knieendoprothesen wird ein Zusammenhang zwischen Leistungsmenge und Behandlungsergebnis vermutet. Ein entsprechender Zusammenhang konnte in der Literatur jedoch bisher nicht belegt werden. Nur wenige Studien untersuchten den Wechseleingriff in der Knieendoprothetik isoliert. Diese unterscheiden sich u. a. in den gewählten Outcome-Parametern und der Größe der verglichenen Volumengruppen.

Die Auswertung von Daten der externen stationären Qualitätssicherung (QS-Daten) kann ergänzend genutzt werden und Hinweise auf mögliche Volumen-Outcome-Beziehungen liefern. Die vorliegende Sonderauswertung untersucht QS-Daten hinsichtlich eines Zusammenhangs zwischen Leistungsmenge und den Ergebnisindikatoren bei Wechseleingriffen von Knieendoprothesen.

Auftrag

Gegenstand der Beauftragung vom 16. Mai 2019 ist die Analyse eines Zusammenhangs zwischen der Zahl der Wechseleingriffe und dem Outcome ausgewählter Ergebnisindikatoren aus dem bis zum Jahr 2014 bestehenden Leistungsbereich „Knie-Endoprothesenwechsel und -komponentenwechsel“. Dabei sollen gemäß Beauftragung folgende, hier gekürzt wiedergegebene Indikatoren aus dem Erfassungsjahr 2014 einbezogen werden:

- QI 51044: Gehunfähigkeit bei Entlassung
- QI 2220: Gefäßläsion/Nervenschaden
- QI 51049: Frakturen
- QI 51874: Postoperative Wundinfektionen ohne präoperative Infektzeichen
- QI 51054: Wundhämatome/Nachblutungen
- QI 51059: Allgemeine postoperative Komplikationen
- QI 51064: Reoperationen aufgrund von Komplikationen
- QI 51069: Todesfälle

Zur Analyse sollten dabei die QS-Daten des Erfassungsjahres 2014 herangezogen werden, wobei eine Erweiterung der Datenbasis grundsätzlich möglich war. Aus Gründen, die in Kapitel 3 erörtert werden, wurde davon allerdings abgesehen.

Methodisches Vorgehen

Zur Analyse potentieller Zusammenhänge zwischen Fallzahl und Patienten-Outcome wird für jeden Indikator jeweils ein sogenanntes *Generalisiertes additives gemischtes Modell* als Inferenzmodell verwendet (vgl. u. a. Wood 2006). Dabei wird ein logistischer Regressionsansatz verwendet, der die bedingte Wahrscheinlichkeit für das Auftreten des jeweils interessierenden Ereignisses (Outcomes) eines Patienten oder einer Patientin in Abhängigkeit mehrerer Einflussgrößen, insbesondere der Fallzahl, modelliert. Als Einflussgrößen fließen von der Fallzahl unabhängige leistungserbringerspezifische sog. Random-Intercepts ein, die die Modellierung einer von der Fallzahl unabhängigen *Grundkompetenz* des Leistungserbringers im jeweils betrachteten Qualitätsaspekt erlauben. Des Weiteren gehen patientenspezifische Risikoscores aus der Risikoadjustierung von AQUA (2015b) in die Modellierung ein, wie sie für die Bundesauswertung zum Erfassungsjahr 2014 verwendet wurden. Der Einfluss der Fallzahl auf das Behandlungsergebnis wird durch eine stetige Funktion $f(n_i)$ der Fallzahl modelliert, die den Zusammenhang zwischen der Wahrscheinlichkeit für das Auftreten des im jeweiligen Indikator interessierenden Ereignisses (unter Festhalten der anderen Größen im Modell) und der Fallzahl beschreibt.

Der Fallzahleffekt *f* wird mithilfe eines penalisierten Spline-Ansatzes als kontinuierliche Funktion der Fallzahl approximiert. Die statistische Bewertung der Ergebnisse erfolgt anhand von simultanen Konfidenzbändern für f zum Konfidenzniveau 95 % bzw. mit einem zu diesen Konfidenzbändern zugehörigen statistischen Test der Hypothese eines von Null verschiedenen Fallzahl-Effekts gegen die (Null-)Hypothese eines nicht existierenden (trivialen) Fallzahl-Effekts zum Signifikanzniveau α = 5 %.

Basierend auf dem oben beschriebenen Ansatz werden für jeden der betrachteten acht Qualitätsindikatoren die Inferenzmodelle geschätzt. Insbesondere ergibt sich eine Schätzung $\hat{f}$ des Fallzahleffekts, was die Berechnung von Odds-Ratios für bestimmte Fallzahlkonstellationen ermöglicht. Außerdem erfolgt eine Schätzung für die Variabilität der leistungserbringerspezifischen Grundkompetenzen (Random-Intercepts), die die nicht auf die Fallzahl zurückzuführende Heterogenität in der Kompetenz der Leistungserbringer quantifiziert. Die geschätzte Variabilität in der Grundkompetenz der Leistungserbringer kann mithilfe sog. Median-Odds-Ratios (vgl. Gleichung (2) in Abschnitt 4.1.3) direkt mit den Odds-Ratios für bestimmte Fallzahlkonstellationen verglichen werden. Dies ermöglicht eine Einordnung der Größenordnung des Fallzahl-Effektes gegenüber der Größenordnung *nicht* fallzahlabhängiger Unterschiede in der Kompetenz der Leistungserbringer.

Ergebnisse

Die Auswertung der aus den statistischen Tests gewonnenen *p*-Werte beziehungsweise die Betrachtung grafischer Darstellungen des geschätzten Volumen-Outcome-Zusammenhangs mit geeigneter Visualisierung von statistischer Unsicherheit führt zu folgenden Ergebnissen:

Bei drei der acht Qualitätsindikatoren zeigt sich ein statistisch signifikanter Fallzahleffekt. Diese sind

- QI 51054: Wundhämatome/Nachblutungen,

- QI 51064: Reoperationen aufgrund von Komplikationen,
- QI 51069: Todesfälle.

In allen drei Fällen hat die geschätzte Funktion $\hat{f}$ einen monoton fallenden Verlauf, d. h. mit steigender Fallzahl und sonst gleichen Modellparametern nimmt die geschätzte Wahrscheinlichkeit für das Eintreten des interessierenden Ereignisses ab. Bei den übrigen fünf Qualitätsindikatoren ist die Schätzunsicherheit des Fallzahleffektes so hoch, dass keine statistisch signifikanten Fallzahleffekte festgestellt werden konnten. Innerhalb der Spannweite der Konfidenzbänder, die das Spektrum mit den Daten „in Einklang zu bringender" Fallzahleffekte abbilden, könnten für diese übrigen fünf Indikatoren sowohl steigende als auch fallende und sogar nicht monotone Verläufe des Fallzahleffektes möglich sein.

Entsprechend der geringen Prävalenz der interessierenden Ereignisse variiert der geschätzte Fallzahleffekt für alle acht Indikatoren allerdings jeweils nur innerhalb eines kleinen Wertebereichs, d. h. der Unterschied der geschätzten Wahrscheinlichkeiten zwischen Leistungserbringern niedriger Fallzahl und Leistungserbringern hoher Fallzahl fällt, bei sonst gleichen Modellparametern, sehr klein aus. Die geschätzte Heterogenität in der fallzahlunabhängigen Grundkompetenz der Leistungserbringer fällt demgegenüber deutlich höher aus.

Fazit

Die vorliegende Analyse bietet allenfalls Hinweise für einen möglichen, versorgungsrelevanten Zusammenhang zwischen Volumen und Outcome eines Standortes. Zwar kann für drei der acht untersuchten Indikatoren ein statistisch signifikanter Zusammenhang zwischen Leistungsmenge und Behandlungsergebnis festgestellt werden, allerdings ist dieser Zusammenhang gegenüber der Variabilität in der fallzahlunabhängigen Grundkompetenz der Leistungserbringer und in Anbetracht der insgesamt sehr niedrigen Prävalenz der betrachteten Outcome-Parameter einzuordnen. Die Analyse unterliegt zudem einer Vielzahl von Limitationen (vgl. Kapitel 6, Seite 49). Diese schließen fehlende patientenseitige Informationen (z. B. Indikation für den Wechseleingriff) ebenso mit ein, wie nicht bekannte, jedoch relevante Strukturmerkmale aufseiten der Leistungserbringer. Darüber hinaus wurde jeder Ergebnisindikator separat analysiert, eine umfassende Beantwortung der Fragestellung würde eine Wichtung der einzelnen Indikatoren hinsichtlich ihrer klinischen Relevanz voraussetzen. Entscheidend ist zudem die Begrenzung der Datenbasis auf das Eintreten intra-hospitaler Ereignisse. Es muss von einer hohen Anzahl klinisch relevanter Ereignisse im poststationären Verlauf ausgegangen werden, die mit vorliegender Analyse nicht erfasst werden kann. Aus medizinischer Sicht verbirgt sich hinter dem Begriff des „Knie-Endoprothesenwechsels bzw. -komponentenwechsels" ein heterogenes Feld an Indikationen und Eingriffen, deren fachliches Anforderungsprofil erheblich divergiert. Die sich daraus ergebende Heterogenität erschwert eine einheitliche Interpretation der Ergebnisse. Auf einen klinisch relevanten Unterschied in der Versorgungsqualität zwischen Standorten allein anhand ihrer Fallzahl kann somit auf Basis der vorliegenden Daten der externen stationären Qualitätssicherung nicht geschlossen werden.

Um die Ausgangshypothese, dass die fachlich anspruchsvollen Knieendoprothesenwechsel einem Volumen-Outcome-Zusammenhang unterliegen, zu überprüfen, sollten zukünftige Untersuchungen das Operationsverfahren präziser eingrenzen. Gleichzeitig sollte die Datenbasis durch Hinzunahme weiterer Datenquellen (z. B. Sozialdaten) erweitert werden, um poststationäre Komplikationen erfassen zu können. Generell zeigt sich somit, dass die Erforschung von Volumen-Outcome-Zusammenhängen eine grundlegende Auseinandersetzung mit der Frage erfordert, welche methodischen Standards an den Nachweis solcher Zusammenhänge gestellt werden sollen, um evidenzbasierte gesundheitspolitische Schlussfolgerungen ziehen zu können.

1 Hintergrund

Analysen zu Zusammenhängen zwischen erbrachter Leistungsmenge und Behandlungsergebnis datieren bis in die 1970er Jahre zurück. Zum Beispiel konnten Luft et al. (1979) eine Reduktion der Sterblichkeit bei Herz-, Gefäß- und Prostata-Operationen mit steigender Fallzahl des Krankenhauses feststellen. Diese „Volumen-Outcome-Hypothese“ wird bis heute kontrovers diskutiert. 2002 erreichte die wissenschaftliche Diskussion mit Einführung des Fallpauschalengesetzes die Gesundheitspolitik: Der Gesetzgeber beauftragte den Gemeinsamen Bundesausschuss (G-BA) mit der Erstellung eines Katalogs „planbarer Leistungen, bei denen die Qualität des Behandlungsergebnisses von der Menge der erbrachten Leistungen abhängig ist“ (§ 136b Abs. 1 Satz 1 Nr. 2 SGB V).

Ausgangsgedanke des zugrunde liegenden Auftrages ist die Annahme, dass bei bestimmten Leistungen eine höhere Fallzahl des Krankenhauses mit einer besseren Versorgungsqualität einhergeht. Als ursächlich für diesen Zusammenhang werden verschiedene Faktoren diskutiert. So wird bei sogenannten „high-volume“-Krankenhäusern ein verbessertes Komplikationsmanagement angenommen (Ghaferi et al. 2009). Die im Vergleich zum größeren Standort geringere Verfügbarkeit von personellen und strukturellen Ressourcen bei kleineren Krankenhäusern kann im Einzelfall eine frühzeitige Diagnostik und Behandlung auftretender Komplikationen verzögern („failure to rescue“). Zudem kann in Krankenhäusern mit hohen Fallzahlen ein höherer Grad an Spezialisierung bestehen (Bachmann et al. 2002). Dies basiert möglicherweise auf einem „Trainingseffekt“ i. S. einer verbesserten Behandlungstechnik der Operateurinnen und Operateure oder eingespielter Prozesse in der ärztlichen, pflegerischen und therapeutischen Nachbehandlung der Patientinnen und Patienten („practice-makes-perfect“) (Luft et al. 1987, Schräder et al. 2007). Letzteres wird auch gefördert durch eine möglicherweise häufigere Ausarbeitung und Anwendung klinischer Behandlungspfade in Krankenhäusern mit größeren Fallzahlen (Bachmann et al. 2002). Dem gegenüber steht die „selective-referral“-Hypothese (Luft et al. 1987, Schräder et al. 2007). Hier wird die Kausalkette umgedreht und die hohe Qualität eines Standortes als ursächlich für eine hohe Leistungsmenge angesehen. So könnte die gute Qualität eines Standortes mit einem guten Ruf einhergehen und so zu einer hohen Anzahl an Patientinnen und Patienten führen. Unabhängig davon können regionale Standortvorteile hohe Fallzahlen generieren. Als Beispiel sei hier die Nähe zu einem Skigebiet genannt, die infolge von vermehrt auftretenden Skiunfällen die Fallzahlen in den umliegenden Krankenhäusern erhöht.

Gegenstand von Untersuchungen zur Beziehung zwischen der Fallzahl eines Krankenhauses und der Ergebnisqualität sind zumeist chirurgische Leistungsbereiche (Pieper et al. 2013). Unter Annahme der im vorherigen Abschnitt formulierten Prämisse wird insbesondere für fachlich anspruchsvolle Eingriffe ein Zusammenhang zwischen Leistungsmenge und Behandlungsergebnis vermutet (Kizer 2003). Aus chirurgischer Sicht stellt sich ein Wechseleingriff im Vergleich zur Primärimplantation einer Knieendoprothese deutlich komplexer dar: Eine detaillierte Planung der Wechseloperation ist essentiell und muss neben der Indikation (z. B. septische oder aseptische Lockerung) auch das Primärimplantat (z. B. unikondyläre Schlittenprothese oder bikondylä-

ren Prothese), die Einbautechnik (zementiert oder unzementiert) sowie patientenseitige Risikofaktoren (z. B. Komorbiditäten und Knochensubstanz) berücksichtigen. Von einem Komponentenwechsel spricht man, wenn lediglich Teile des Implantats (z. B. das Inlay) gewechselt werden. Je nach vorliegender Situation kann die Operation im Rahmen eines ein- oder zweizeitigen Wechsels (Entnahme der alten und Implantation der neuen Endoprothese in zwei Eingriffen mit zeitlichem Abstand) durchgeführt werden.

Die dargestellte Komplexität des Wechseleingriffs deutet bereits mögliche Schwierigkeiten in der Analyse einer Volumen-Outcome-Beziehung an. Da die Mehrzahl der Studien Primär- und Sekundäreingriffe zusammenfasst, stehen nur wenige Studien zu Verfügung, die Zusammenhänge isoliert für Knieendoprothesenwechsel berichten. Zwei US-amerikanische Studien konnten im Rahmen einer orientierenden Literatursuche identifiziert werden. Diese untersuchen auf Basis von Abrechnungsdaten mögliche Zusammenhänge zwischen Fallzahl und Behandlungsergebnis für Revisionseingriffe in der Knieendoprothetik (Feinglass et al. 2004, Taylor et al. 1997).

Taylor et al. (1997) nutzten für ihre Analyse die „National Medicare Database“ aus den Jahren 1992 bis 1994. Annähernd 23.000 Knieendoprothesen-Wechsel wurden dafür eingeschlossen und in drei Volumengruppen unterteilt. Diese bezogen sich auf die Leistungsmenge des gesamten Krankenhauses und lagen bei weniger als 25 Eingriffen, zwischen 25 und 199 Eingriffen und mehr als 199 Eingriffen pro Jahr. Die Krankenhausmortalität und die 30-Tages-Mortalität wurden i. S. eines Behandlungsergebnisses als Outcome-Parameter definiert. Als mögliche weitere Einflussfaktoren wurden das Alter und das Geschlecht der Patientinnen und Patienten berücksichtigt. Ein statistisch signifikanter Zusammenhang zwischen Leistungsmenge des Krankenhauses und der Mortalität wurde in der Studie nicht festgestellt.

Ebenfalls auf US-amerikanischen Abrechnungsdaten beruht eine weitere Analyse (Feinglass et al. 2004). Im Untersuchungszeitraum 1993 bis 1999 wurden annähernd 3.000 Patientinnen und Patienten mit einem Knieendoprothesen-Wechsel in die Untersuchung eingeschlossen. Als primäres Outcome wurde das Auftreten einer Komplikation definiert. Die Autoren wählten ebenfalls drei Volumengruppen zum Vergleich, setzten jedoch andere Schwellenwerte („Low-volume“: 7 bis 48 Fälle; „medium-volume“ 49 bis 98 Fälle und „high-volume“ mehr als 98 Fälle). Auch in dieser Studie konnte keine signifikante Assoziation zwischen den gewählten Volumengruppen und dem Auftreten einer Komplikation nachgewiesen werden.

Weder für die Mortalität noch für die Komplikationsrate konnte in der Literatur ein Zusammenhang zwischen Leistungsmenge und Behandlungsergebnis dargestellt werden. Generell beeinflussen gewählte Größe, Gruppengrenze und Anzahl der untersuchten Volumengruppen das Ergebnis einer Volumen-Outcome-Analyse entscheidend (Abschnitt 4.4), sodass ein Rückschluss auf klinisch relevante Zusammenhänge in entsprechenden Studien kritisch zu bewerten ist (Bender und Grouven 2006). Alternativ ist die Analyse der Leistungsmenge i. S. einer kontinuierlichen Variable in Betracht zu ziehen (vgl. Abschnitt 4.2). Eine entsprechende Auswertung zu Wechseleingriffen in der Knieendoprothetik ist dem IQTIG jedoch nicht bekannt.

In der Zusammenschau erschweren verschiedene Faktoren die Beantwortung der vordergründig einfachen Fragestellung nach klinisch relevanten Volumen-Outcome-Zusammenhängen. Aus der medizinischen Komplexität des Eingriffs ergibt sich ein heterogenes Feld an Eingriffen, die

als Knieendoprothesenwechsel zusammengefasst werden können. Zudem spielen die gewählten statistischen Methoden eine entscheidende Rolle in der Beurteilung möglicher Rückschlüsse auf die Versorgungssituation. Die vorliegende Sonderauswertung untersucht Daten der externen stationären Qualitätssicherung auf Hinweise zu Zusammenhängen zwischen der erbrachten Leistungsmenge und der Ergebnisqualität.

2 Auftrag und Auftragsverständnis

Gegenstand der Beauftragung vom 16. Mai 2019 ist eine Sonderauswertung zur Darstellung der Beziehung zwischen Leistungsmenge und Behandlungsergebnis bei Knieendoprothesen-Wechseleingriffen anhand von Daten der externen stationären Qualitätssicherung (G-BA 2019). Datengrundlage sind die vom AQUA-Institut erhobenen Daten des damaligen Leistungsbereichs „Knie-Endoprothesenwechsel- und komponentenwechsel“ aus dem Erfassungsjahr 2014 (AQUA 2015a). Als Outcome-Parameter sind gemäß Punkt I.2 der Beauftragung die folgenden, hier verkürzt dargestellten Ergebnisindikatoren definiert (AQUA 2015b):

- QI-ID 51044: Gehunfähigkeit bei Entlassung
- QI-ID 2220: Gefäßläsion/Nervenschaden
- QI-ID 51049: Frakturen
- QI-ID 51874: Postoperative Wundinfektionen ohne präoperative Infektzeichen
- QI-ID 51054: Wundhämatome/Nachblutungen
- QI-ID 51059: Allgemeine postoperative Komplikationen
- QI-ID 51064: Reoperationen aufgrund von Komplikationen
- QI-ID 51069: Todesfälle

Potentielle Zusammenhänge sollen gemäß Punkt I.1. des Auftrags anhand geeigneter Volumengruppen untersucht werden. Sofern durch Einbezug weiterer Erfassungsjahre eine höhere Aussagekraft erwartet wird, kann dies schriftlich beantragt werden.

Die Fragestellung, die verfügbare Datenbasis sowie die vorgeschlagenen Methoden zur Bearbeitung des Auftrags werden vom IQTIG im vorliegenden Bericht geprüft und diskutiert. Direkt und ohne Anpassungen übernommen wurden dabei die Definitionen und Rechenregeln der Qualitätsindikatoren gemäß AQUA (2015b) einschließlich der Risikoadjustierungsmodelle.

Hervorzuheben ist, dass dies die erstmalige Beauftragung des IQTIG zur Analyse potentieller Volumen-Outcome-Zusammenhänge ist. Entsprechend kommt der Darstellung und Diskussion der gewählten Methodik in vorliegendem Bericht eine besondere Bedeutung zu. Da die Ergebnisse von Volumen-Outcome-Analysen erheblich von der Anzahl, Größe und Begrenzung der gewählten Volumengruppen abhängen, ist aus Sicht des IQTIG die unter Punkt I.1. der Beauftragung gewünschte Bildung von Volumengruppen und -grenzen nicht sinnvoll (Grouven et al. 2008, Wetzel 2006, IQWIG 2019). Die vorliegende Analyse basiert daher auf der Annahme, dass allenfalls eine stetige Beziehung zwischen Fallzahl und Indikatorergebnis besteht (näheres in Abschnitt 4.4 bzw. auf Seite 32).

Es ist zu unterstreichen, dass eine Aussage über die Kausalität dargestellter Fallzahl-Ergebnis-Zusammenhänge anhand der QS-Daten nicht möglich ist und die Fallzahl auch nicht als mögliche kausale Ursache für bessere oder schlechtere Behandlungsqualität interpretiert werden sollte. Vielmehr ist die Fallzahl als Surrogat-Parameter aufzufassen, der potentiell mit der Kompetenz des Leistungserbringers korreliert (Wetzel 2006). Die vermutete höhere Kompetenz bei „high-volume“-Krankenhäusern könnte sich, wie oben dargestellt (vgl. Kapitel 1), aus besseren struk-

turellen, personellen und prozeduralen Voraussetzungen und Möglichkeiten ergeben, die wiederum in einem direkten kausalen Zusammenhang zum Behandlungsergebnis stehen können. Der Surrogat-Parameter „Fallzahl“ kann somit höchstens mittelbar über die *Kompetenz des Leistungserbringers* einen Einfluss auf das Behandlungsergebnis von Patientinnen und Patienten haben (vgl. Abschnitt 4.1). Die übergeordnete Fragestellung des Auftrags wird daher wie folgt verstanden: Besteht ein Zusammenhang zwischen der Kompetenz eines Leistungserbringers in einem betrachteten Qualitätsaspekt (Ergebnisindikator) und der Anzahl an durchgeführten Knieendoprothesen-Wechseln (Fallzahl) im betrachteten Erfassungsjahr?

3 Beschreibung der Datenbasis und der Indikatoren

Als Datengrundlage für die Analyse von Fallzahl-Ergebnis-Zusammenhängen bei Knieendoprothesenwechseln sieht der Auftrag (G-BA 2019) die Daten des Qualitätssicherungsverfahrens „Knie-Endoprothesenwechsel und -komponentenwechsel“ aus dem Erfassungsjahr 2014 (vgl. AQUA 2015b) vor.

Insgesamt wurden dabei 17.658 Wechselprozeduren in 17.551 Behandlungsfällen dokumentiert, die in 1.058 Krankenhausstandorten erbracht wurden (AQUA 2015a).

Verteilung der Fallzahlen

Als Grundlage für die Analysen wird zunächst dargestellt, welche Standortfallzahlen mit welcher Häufigkeit auftreten. Abbildung 1 stellt die absolute Häufigkeit der Standorte differenziert nach Fallzahl in einem Histogramm dar. Mit n_i wird dabei im gesamten Bericht die Fallzahl eines Standortes i bezeichnet, vgl. das Glossar mathematischer Notationen. Es treten Fallzahlen zwischen 1 und 495 auf. Aufgrund des breiten Spektrums der auftretenden Fallzahl und der Konzentration der Standorte im Bereich niedriger Fallzahlen von 1 bis ca. 20 wurde eine logarithmische Achsenskalierung für diese und alle weiteren fallzahlspezifischen Darstellungen gewählt.

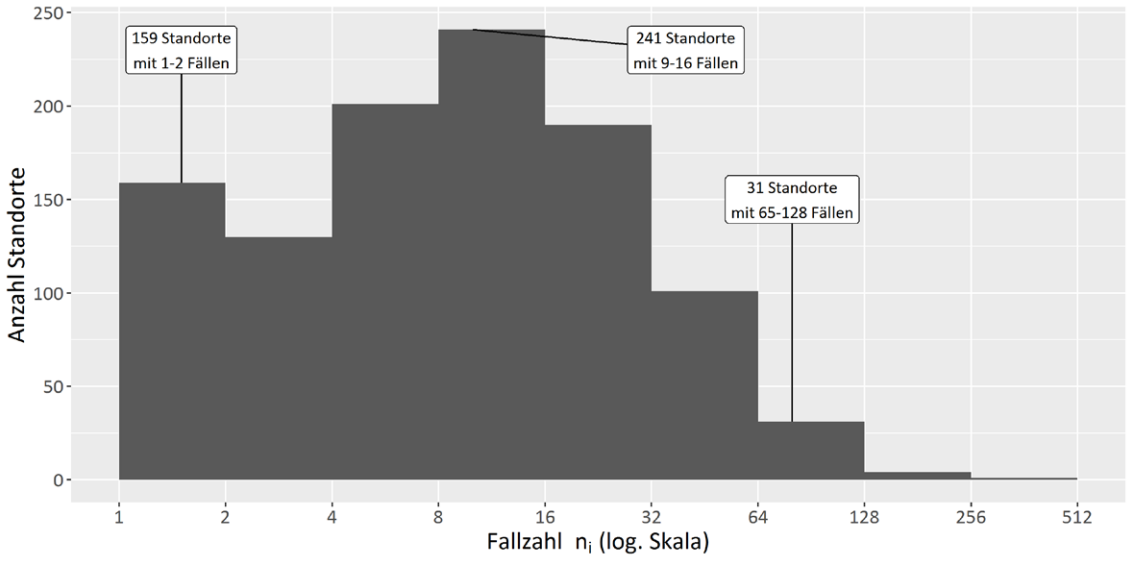

Abbildung 1: Verteilung der Standorte differenziert nach Fallzahl (x-Achse logarithmisch skaliert)

Zu erkennen ist, dass die meisten Standorte im Erfassungsjahr 2014 eine Fallzahl von 5 bis 32 Wechselprozeduren hatten. Nur wenige Standorte führten mehr als 128 Wechselprozeduren im Erfassungsjahr 2014 durch. Auffällig ist auch eine große Anzahl von Standorten mit nur 1–2 durchgeführten Wechselprozeduren (159 Standorte). Abbildung 2 zeigt, dass diese Standorte

bezogen auf die Gesamtversorgung nur einen sehr kleinen Teil an allen Knieendoprothesenwechseln durchführen. Dargestellt ist der Anteil an allen Knieendoprothesenwechseln differenziert nach Fallzahl. Die Mehrheit der Patientinnen und Patienten werden somit in Standorten von ca. 9 bis ca. 128 Fällen versorgt.

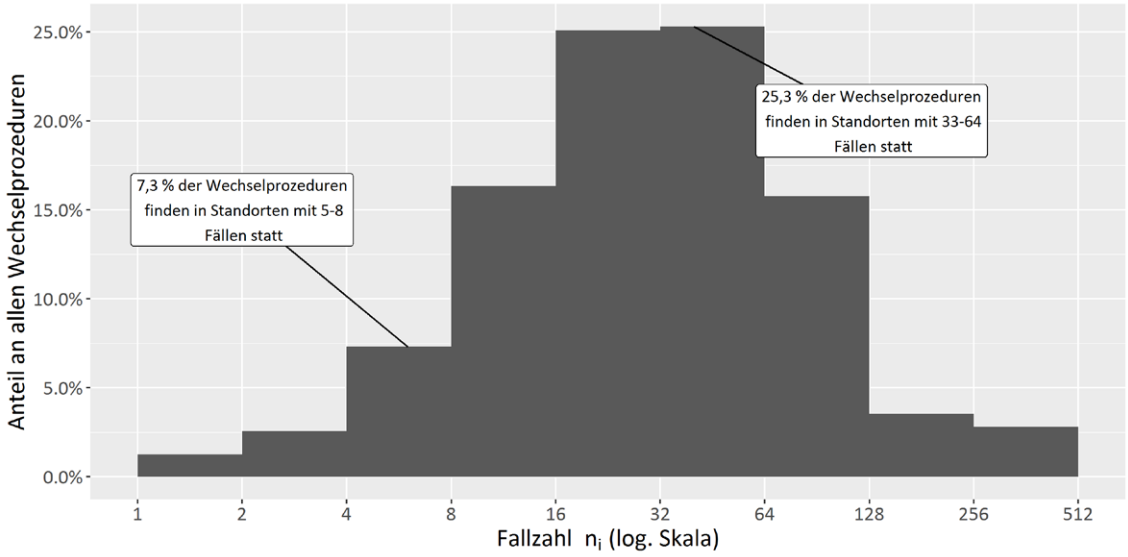

Abbildung 2: Anteil an der Gesamtversorgung von Knieendoprothesenwechseln differenziert nach Fallzahl (x-Achse logarithmisch skaliert)

Qualitätsindikatoren und Prävalenz der Outcome-Parameter

Der Auftrag sieht die Analyse von Volumen-Outcome-Beziehungen bei den in Tabelle 1 genannten Qualitätsindikatoren vor (G-BA 2019):

Tabelle 1: Zur Analyse vorgesehene Ergebnisindikatoren des Leistungsbereichs „Knie-Endoprothesenwechsel und -komponentenwechsel“ aus dem Erfassungsjahr 2014.

QI-ID	Kurzbezeichnung	Risikoadjustierung vorhanden
51044	Gehunfähigkeit bei Entlassung	ja
51049	Frakturen	ja
51054	Wundhämatome/Nachblutungen	ja
51059	Allgemeine postoperative Komplikationen	ja
51064	Reoperationen aufgrund von Komplikationen	ja
51069	Todesfälle	ja
2220	Gefäßläsionen/Nervenschäden	nein

QI-ID	Kurzbezeichnung	Risikoadjustierung vorhanden
51874	Postoperative Wundinfektionen ohne präoperative Infektzeichen	nein

Die Indikatoren wurden unverändert gemäß ihrer Rechenregeldefinitionen inklusive Risikoadjustierung aus (AQUA 2015a) übernommen.

Tabelle 2 zeigt die Ereignisprävalenz für alle acht Indikatoren, d. h. den Anteil an interessierenden Ereignissen unter allen Prozeduren/Behandlungsfällen, die in den jeweiligen Qualitätsindikator eingehen. Außerdem wird der Anteil an Standorten dargestellt, bei denen jeweils kein interessierendes Ereignis auftrat.

Tabelle 2: Ereignisprävalenz und Anteil an Standorten ohne Vorkommen interessierender Ereignisse für die acht zur Analyse vorgesehenen Ergebnisindikatoren.

Qualitätsindikator	Ereignisprävalenz (Zähler/Nenner, Anteil)	Anteil an Standorten ohne Vorkommen eines interessierenden Ereignisses
Gehunfähigkeit bei Entlassung	118 / 17.482 (0,67 %)	90,5 %
Frakturen	68 / 17.658 (0,39 %)	94,6 %
Wundhämatome/ Nachblutungen	383 / 17.658 (2,17 %)	75,5 %
Allgemeine postoperative Komplikationen	339 / 17.551 (1,93 %)	78,4 %
Reoperationen aufgrund von Komplikationen	657 / 17.658 (3,72 %)	65,3 %
Todesfälle	69 / 17.551 (0,39 %)	94,0 %
Gefäßläsionen/ Nervenschäden	28 / 17.658 (0,16 %)	97,4 %
Postoperative Wundinfektionen ohne präoperative Infektzeichen	45 / 8.523 (0,53 %)	95,4 %

An der teilweise sehr niedrigen Ereignisprävalenz zeigt sich, dass die Versorgungsqualität bei vielen der betrachteten Qualitätsindikatoren insgesamt auf einem hohen Niveau war. Eine Sonderrolle nehmen die beiden Qualitätsindikatoren „Reoperationen aufgrund von Komplikationen“ und „Allgemeine postoperative Komplikationen“ ein. Reoperationen treten als Folge schwerer spezifischer behandlungsbedürftiger Komplikationen auf. Somit gehen in diesen Qualitätsindikator Fälle ein, die auch mindestens in einem der anderen Qualitätsindikatoren auftraten. Insbesondere hat der Qualitätsindikator große Überschneidungen mit den Qualitätsindikatoren „Frakturen“, „Wundhämatome/Nachblutungen“, „Gefäßläsionen/Nervenschäden“ und „Postoperative Wundinfektionen“. Der Qualitätsindikator hat somit zusammenfassenden Charakter für derartige spezifische behandlungsbedürftige Komplikationen.

Während der Qualitätsindikator „Reoperationen aufgrund von Komplikationen“ eine implizite Zusammenfassung von spezifischen Komplikationen darstellt, werden im Qualitätsindikator „Allgemeine postoperative Komplikationen“ explizit mehrere Komplikationen zusammengefasst: Pneumonien, kardiovaskuläre Komplikationen, tiefe Bein-/Beckenvenenthrombosen und Lungenembolien.

Während die niedrige Prävalenz von Komplikationen für die Versorgungsqualität ein gutes Zeichen ist, ergibt sich daraus für die statistische Analyse von Fallzahl-Ergebnis-Zusammenhängen, dass alle modellbasierten Schätzungen einer großen statistischen Unsicherheit unterliegen. Es ist somit schon a priori ersichtlich, dass potentielle Fallzahl-Ergebnis-Zusammenhänge sehr ausgeprägt sein müssen, um als statistisch signifikante Ergebnisse messbar zu sein. Um diesen Gedanken zu formalisieren, könnte man für ein einfaches Analysemodell Abschätzungen berechnen, welche Effektstärke mit der vorhandenen Datenbasis überhaupt messbar gewesen wäre. Da die Datengrundlage im Rahmen des Auftrags nicht sinnvoll erweiterbar ist (vgl. nächsten Absatz), wurde davon abgesehen.

Erweiterung der Datenbasis auf mehrere Erfassungsjahre

Der Auftrag sieht die Möglichkeit vor, die Datenbasis auf die Erfassungsjahre 2013 und 2012 oder neuere Daten auszuweiten, um die statistische Aussagekraft der Analysen zu stärken. Auf die Möglichkeit, Daten vor dem Erfassungsjahr 2014 einzubeziehen, wurde jedoch nicht zurückgegriffen, da es zwischen den Erfassungsjahren 2013 und 2014 eine Umstellung in der Auswertung gab: Bis einschließlich zum Erfassungsjahr 2013 wurde ein Leistungserbringer als Auswertungseinheit über das Institutionskennzeichen („IK-Nummer“) definiert. Seit dem Erfassungsjahr 2014 wird allerdings der entlassende Standort als Auswertungseinheit verwendet, somit ist die Auswertungseinheit spezifischer. Für die Frage nach einem Fallzahl-Ergebnis-Zusammenhang und insbesondere dessen Interpretation ist die genaue Definition der Auswertungseinheit sehr wichtig. Interpretiert man die Fallzahl als Surrogat-Parameter für bestimmte Strukturmerkmale des Leistungserbringers (Erfahrung des Operationsteams, OP-Ausstattung, etc.), so ist diese Interpretation der Fallzahl nur dann valide, wenn die relevanten Strukturmerkmale innerhalb einer Auswertungseinheit konstant sind, was auch in den hier verwendeten Analysemodellen angenommen werden muss (vgl. auch Kapitel 4 und Kapitel 6). Tatsächlich ist diese Annahme bereits auf der feineren Ebene der entlassenden Standorte relativ stark: Beispielsweise für Qualitätsindikatoren, die intraoperative Komplikationen erfassen, wären noch kleinere Auswertungseinheiten zu bevorzugen, weil davon auszugehen ist, dass die Erfahrung und die Zusammenstellung des jeweiligen OP-Teams wichtige Einflussgrößen für das Behandlungsergebnis sind und solche Teams auch innerhalb eines Standortes große Unterschiede aufweisen können. Vor diesem Hintergrund erscheint eine Verallgemeinerung über die Ebene der entlassenden Standorte hinaus umso weniger sinnvoll.

Eine Verbreiterung der Datenbasis ist somit nur sinnvoll, wenn die Datenbasis auf neuere Daten ausgeweitet wird, die wenigstens auch die Standortebene mit ausweisen. Wegen der Zusammenführung der Leistungsbereiche „Knie-Totalendoprothesen-Erstimplantation“ (17/5) und „Knie-Endoprothesenwechsel und -komponentenwechsel“ (17/7) zum QS-Verfahren *Knieendoprothesenversorgung* (*KEP*) ab dem Erfassungsjahr 2015 ist auch das nur sinnvoll, wenn

man die Analyse vollständig auf das aktuelle QS-Verfahren, die aktuellen Qualitätsindikatoren und neuere Daten (Erfassungsjahre 2015 bis 2018) bezieht. Da der Auftrag explizit die Verwendung der Daten aus dem Erfassungsjahr 2014 vorsieht, wurde davon abgesehen.

4 Methodik

Zur Analyse potentieller Zusammenhänge zwischen Fallzahl und Patienten-Outcome wird für jeden Qualitätsindikator ein sogenanntes *Generalisiertes additives gemischtes Modell* verwendet (vgl. u. a. Wood 2006):

$$\mathrm{Logit}\big(\mathrm{P}\big(y_{ij}=1\big|\eta_{ij},n_i,u_i\big)\big)=\beta_0+\eta_{ij}+f(n_i)+u_i. \quad (1)$$

Dabei wird ein logistischer Regressionsansatz verwendet, der die bedingte Wahrscheinlichkeit $\mathrm{P}\big(y_{ij}=1\big|\eta_{ij},n_i,u_i\big)$ für das Auftreten des jeweiligen interessierenden Ereignisses $y_{ij}=1$ eines Patienten oder einer Patientin j bei Leistungserbringer i in Abhängigkeit mehrerer Einflussgrößen, insbesondere der Fallzahl n_i, modelliert. Hierbei ist β_0 ein globaler Intercept-Parameter, Abweichungen davon werden durch von der Fallzahl unabhängige leistungserbringerspezifische Random-Intercepts $u_i \overset{\text{u.i.v.}}{\sim} \mathcal{N}(0,\tau^2)$ modelliert, $f(n_i)$ ist eine stetige Funktion der Fallzahl, die deren Einfluss auf die Wahrscheinlichkeit für das Eintreten des interessierenden Ereignisses (unter Festhalten der anderen Größen im Modell) modelliert und η_{ij} stellen Logit-transformierte patientenspezifische Risikoscores aus der Risikoadjustierung von (AQUA 2015b) dar.

In Abschnitt 4.1 wird die Modellformel (1) aus einem sog. grafischen Modell hergeleitet und begründet. In Abschnitt wird die genaue Methode zur Schätzung des Fallzahleffekts *f*, ein sog. penalisierter Spline-Ansatz, erläutert. Abschnitt 4.3 erläutert, wie die geschätzten Fallzahl-Ergebnis-Zusammenhänge statistisch beurteilt werden. Schließlich wird in Abschnitt 4.2 die gewählte Methodik in Abgrenzung zu anderen statistischen Methoden, insbesondere Methoden, die die Bildung von Volumengruppen erfordern, diskutiert.

Alle hier und im Folgenden verwendeten mathematischen Symbole und Notations-Konventionen werden auch im Glossar mathematischer Notationen ausgewiesen.

4.1 Herleitung des Regressionsmodells zur Inferenz eines Fallzahl-Ergebnis-Zusammenhangs

Im Folgenden wird das Inferenzmodell (1) zur Bestimmung eines potentiellen Zusammenhangs zwischen Fallzahl und Ergebnisqualität für einen Qualitätsindikator anhand eines sog. grafischen Modells hergeleitet. Allgemein ist ein grafisches Modell eine Visualisierung der Beziehungen unterschiedlicher Größen in einem statistischen Modell (vgl. u. a. Bishop 2006: 359 ff.).

Da die Daten der Qualitätssicherung aus einer retrospektiven Beobachtungsstudie entstammen, ist ein Nachweis einer kausalen Beziehung zwischen Fallzahl und Ergebnisqualität im Allgemeinen nahezu unmöglich, denn im Gegensatz zu einer prospektiven randomisierten Interventionsstudie können potentielle Einflussgrößen auf das Behandlungsergebnis nicht durch ein gezieltes Studiendesign unabhängig voneinander ausgewählt werden. Insbesondere ist für einen Kausalnachweis von Ursache zu Wirkung immer auch ein Nachweis der klaren zeitlichen Abfolge von Ursache und Wirkung erforderlich. Da die Anzahl an Prozeduren eines Erfassungsjahres erst am Ende eines Jahres feststeht, kann logisch somit kein Kausalzusammenhang zum Behandlungsergebnis eines Patienten oder einer Patientin im gleichen Erfassungsjahr bestehen. Vielmehr ist

davon auszugehen, dass die Fallzahl des Krankenhauses mit anderen Einflussgrößen korreliert, beispielsweise mit der Erfahrung der einzelnen Chirurgin bzw. des einzelnen Chirurgen oder mit der Ausstattung des OP-Saals hinsichtlich der Erfordernisse spezialisierter Eingriffe wie Endoprothesen-Wechsel. Die Modellierung von Patienten-Outcomes und derer Einflussgrößen in grafischen Modellen stellt eine Möglichkeit dar, diese komplexen Beziehungen darzustellen. Das grafische Modell dient dabei als Denkhilfe, um präzise formulieren zu können, worin genau der zu messende Zusammenhang zwischen Fallzahl und Ergebnisqualität besteht und verdeutlicht die Rolle der Fallzahl als Surrogat-Parameter für nicht erhobene Leistungserbringermerkmale.

Zunächst wird ein Modell aufgestellt, welches die im Folgenden getroffenen Annahmen über potentielle *kausale* Einflussgrößen auf ein Patienten-Outcome y_{ij} eines Patienten oder einer Patientin j, behandelt durch Leistungserbringer i, beschreibt. Zur Untersuchung eines potentiellen Fallzahl-Ergebnis-Zusammenhangs ist es wichtig zu verstehen, welche Rolle die Fallzahl in diesem Kausalmodell hat. Das kausale grafische Modell wiederum ist die theoretische Grundlage, um das später verwendete Inferenzmodell für einen potentiellen Zusammenhang zwischen Fallzahl und Ergebnisqualität herzuleiten, das letztlich zur statistischen Modellierung des Zusammenhangs verwendet wird.

4.1.1 Kausales grafisches Modell

Das kausale grafische Modell ist in Abbildung 3 dargestellt. Jeder Knoten beschreibt eine Variable bzw. potentielle Einflussgröße und Pfeile beschreiben deren kausale Beziehungen untereinander. Es wird dabei zwischen beobachteten und unbeobachteten Größen unterschieden. Blau eingefärbt sind dabei solche Knoten, die beobachteten Variablen entsprechen. Zunächst besteht das Modell aus zwei Ebenen: Auf oberster Ebene (grüne Box) sind Variablen und Einflussgrößen, die sich von Patientin zu Patientin bzw. von Patient zu Patient unterscheiden können. In der zweiten, unteren Ebene (orange Box) befinden sich Einflussgrößen, die jeweils innerhalb eines Leistungserbringers (LE) konstant sind, d. h. deren Ausprägung für alle Patientinnen und Patienten eines Leistungserbringers gleich ist und somit für die fallübergreifende zugrunde liegende Kompetenz des Leistungserbringers hinsichtlich des im Qualitätsindikator betrachteten Qualitätsaspektes maßgeblich ist.

Wie bei jedem Modell handelt es sich auch beim vorliegenden kausalen grafischen Modell um eine Abstraktion der klinischen Realität, welche Vereinfachungen vornimmt. Beispielsweise werden in diesem Fall keine Interaktionen zwischen Patientenmerkmalen und Leistungserbringerkompetenz abgebildet. Somit werden Phänomene, wie etwa eine besondere Kompetenz bei der Versorgung spezieller Patientengruppen, nicht modelliert. Dieser Aspekt wird auch bei der Diskussion des späteren Inferenzmodells (Abschnitt 4.1.3) und in der Diskussion (Kapitel 6) aufgegriffen.

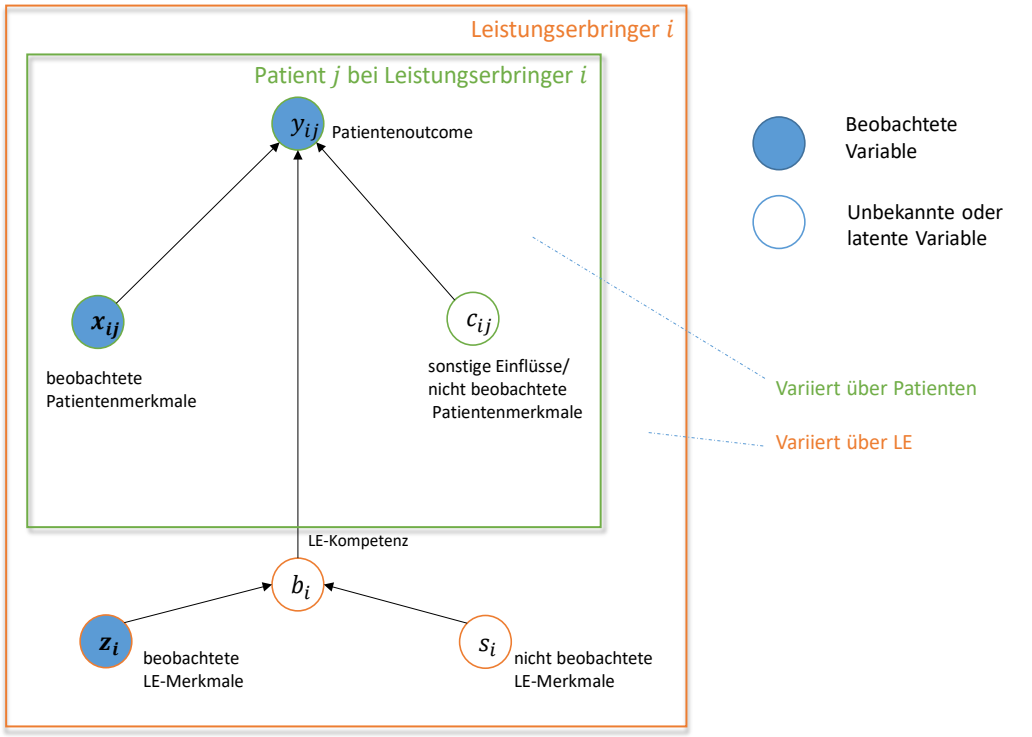

Abbildung 3: Grafisches Modell der kausalen Einflussgrößen auf ein Patienten-Outcome

Das kausale Modell geht davon aus, dass das (binäre) Patienten-Outcome y_{ij} durch drei Gruppen von Einflussgrößen kausal beeinflusst wird (vgl. IQTIG 2019b: 180 ff.): beobachtete Patientenmerkmale $\boldsymbol{x}_{ij} = (x_{ij}^1, \dots, x_{ij}^M)$, die Kompetenz des Leistungserbringers im betrachteten Qualitätsaspekt b_i sowie unbeobachtete sonstige Einflüsse, subsummiert in einem grafischen Knoten c_{ij}.

Der kausale Einfluss der einzelnen Patientenmerkmale hängt dabei nicht nur von der Ausprägung dieser Markmale, sondern auch von deren Relevanz als Risikofaktor für das betrachtete Patienten-Outcome ab. Beispielsweise ist die ASA-Klassifikation ein wichtiger Prädiktor für den Qualitätsindikator „Allgemeine postoperative Komplikationen“ (QI-ID 51059), jedoch nicht relevant für intraoperativ entstandene Frakturen (QI-ID 54049) (vgl. Risikoadjustierung in AQUA 2015b). Diese Relevanz bzw. Stärke der jeweiligen Einflussfaktoren könnte beispielsweise über ein logistisches Regressionsmodell, wie es zur Risikoadjustierung verwendet wird, aus beobachteten Daten geschätzt werden.

Die Kompetenz des Leistungserbringers wiederum wird kausal durch zwei Gruppen von Einflussgrößen beeinflusst: nicht beobachtete Merkmale des Leistungserbringers (z. B. Arbeitsmotivation psychologische Stressfaktoren), subsummiert in einem grafischen Knoten s_i, und beobachtete Merkmale des Leistungserbringers $\boldsymbol{z_i} = \left(z_i^1, \dots, z_i^L\right)$ (z. B. bestimmte Merkmale der OP-Ausstattung, das Vorhandensein einer Blutbank oder die gesammelte Erfahrung der einzelnen OP-Teams in Jahren). Ähnlich wie bei den beobachteten Patientenmerkmalen kann die Relevanz bzw. Stärke der jeweiligen Merkmale für das Patienten-Outcome indikatorspezifisch verschieden sein. Beispielsweise könnte das Vorhandensein einer

Blutbank für die Versorgung operationsbedürftiger Nachblutungen (QI-ID 51054) relevant sein, jedoch nicht zum Entstehen von Gefäßläsionen oder Nervenschäden (QI-ID 2220) beitragen.

Würde man den kausalen Zusammenhang zwischen den Leistungserbringermerkmalen $\boldsymbol{z}_i$ bzw. s_i und Patienten-Outcome genau kennen, so würde sich die Frage nach einem potentiellen Fallzahl-Ergebnis-Zusammenhang erübrigen: Man könnte direkt die Relevanz bestimmter Strukturmerkmale des Leistungserbringers für eine erfolgreiche Behandlung beurteilen und wäre nicht auf den Surrogat-Parameter Fallzahl angewiesen.

4.1.2 Vom Kausalmodell zum Inferenzmodell

Wir betrachten nun den Übergang vom hypothetischen Kausalmodell zum Inferenzmodell, welches in den folgenden Analysen Anwendung findet. Teilt man die potentiellen kausalen Einflussgrößen auf das Patienten-Outcome wie im grafischen Modell Abbildung 3 in patientenseitige Einflüsse $\boldsymbol{x}_{ij}$, Leistungserbringerkompetenz b_i, sowie sonstige Einflüsse c_{ij} auf, so ist klar, dass das Ziel der Volumen-Outcome-Analyse im Kern darin besteht, den Zusammenhang von Fallzahl und Leistungserbringerkompetenz zu quantifizieren, da nur mittelbar über die Kompetenz des Leistungserbringers ein Zusammenhang zwischen Fallzahl und Patienten-Outcome bestehen kann. Dazu ist in Abbildung 4(a) ein Ausschnitt des Kausalmodells dargestellt, welcher nur den Einfluss auf die Kompetenz des Leistungserbringers beschreibt. In Abbildung 4(b) wird dieser Teil des Kausalmodelles um den Knoten der Leistungserbringerfallzahl ergänzt. Es ist davon auszugehen, dass sowohl beobachtete Merkmale des Leistungserbringers als auch die nicht beobachteten Merkmale des Leistungserbringers im Zusammenhang zur Fallzahl stehen. Beispielsweise schafft das Vorhandensein ausreichender OP-Kapazitäten überhaupt erst die Voraussetzung bestimmte Prozedurzahlen pro Jahr durchzuführen. Es wird hier jedoch nicht davon ausgegangen, dass die Fallzahl einen direkten kausalen Einfluss auf die Kompetenz des Leistungserbringers hat. Bedingt auf die beobachteten und nicht beobachteten Leistungserbringermerkmale sind Leistungserbringerkompetenz und Fallzahl unabhängig:

$$b_i|\boldsymbol{z}_i, s_i \perp n_i|\boldsymbol{z}_i, s_i$$

In Abbildung 4(b) wird die Richtung des Pfeils von $\boldsymbol{z}_i$ und s_i zu n_i eingezeichnet. Diese Richtung des Kausalzusammenhangs ist keineswegs eindeutig. Beispielsweise könnte man annehmen, dass sich die Anschaffung einer bestimmten OP-Ausstattung erst ab einer gewissen Fallzahl lohnt und die OP-Ausstattung auf diese Weise von der Fallzahl abhängt. Somit hätte die Fallzahl auch mittelbar einen kausalen Einfluss auf die Behandlungsqualität. Um diese Beziehung eindeutig darzustellen, könnte das grafische Modell um eine zeitliche Komponente erweitert werden, in der man die Fallzahl im betrachteten Erfassungsjahr der Behandlung von der Fallzahl des Leistungserbringers vorangegangener Jahre unterscheidet. Während die aktuelle Fallzahl temporal-kausal keinen Einfluss auf die Behandlungsqualität haben kann, ist es logisch möglich, dass die Fallzahl der letzten Jahre durchaus mittelbaren Einfluss hat. Diese Richtung des Kausalzusammenhangs zwischen $\boldsymbol{z}_i$ und s_i zu n_i ist für die Interpretation und mögliche Konsequenzen

aus der Analyse wichtig[1], spielt allerdings für die weitere Herleitung des Inferenzmodells keine Rolle. Für das Inferenzmodell wird nun in Abbildung 4(c) die Leistungserbringerkompetenz b_i zerlegt in zwei Komponenten:

$$b_i = u_i + v_i$$

Dabei ist u_i der Teil der Leistungserbringerkompetenz, der nicht von der Fallzahl abhängt, d. h.

$$u_i \perp n_i,$$

und v_i der fallzahlabhängige Teil der Leistungserbringerkompetenz. Dabei wird, bedingt auf die Fallzahl, Unabhängigkeit zwischen den beiden Komponenten u_i und v_i angenommen:

$$u_i \perp v_i | n_i$$

Die Variable u_i wird im weiteren Text auch als *Grundkompetenz* des Leistungserbringers bezeichnet. Gestrichelte Linien von $\boldsymbol{z}_i$ und s_i zu u_i deuten an, dass sowohl beobachtete als auch unbeobachtete Leistungserbringermerkmale aus dem Kausalmodell Einfluss auf die fallzahlunabhängige Grundkompetenz u_i haben könnten. Zu beachten ist allerdings, dass für das Inferenzmodell (bedingt durch die Nicht-Verfügbarkeit entsprechender Daten in der QS-Dokumentation) keine Angaben zu spezifischen Leistungserbringermerkmalen zur Verfügung stehen. Für den fallzahlabhängigen Teil v_i wird angenommen, dass er sich als $v_i = f(n_i)$ schreiben lässt, mit einer für alle Leistungserbringer gleichen Funktion f. Außerdem wird angenommen, dass der Knoten v_i nur von der Fallzahl und nicht von anderen Parametern abhängt. Kenntnis der Funktion f zu erlangen ist das Ziel des Inferenzmodelles, welches in Abbildung 4(d) ohne die Knoten (der unbekannten Variablen) des Kausalmodelles dargestellt wird.

[1] Die Richtung des Pfeils entscheidet über die Interpretation des Zusammenhangs zwischen Fallzahl und Ergebnisqualität als „selective referral“ oder „practice makes perfect“ im Sinne von Schräder et al. (2007).

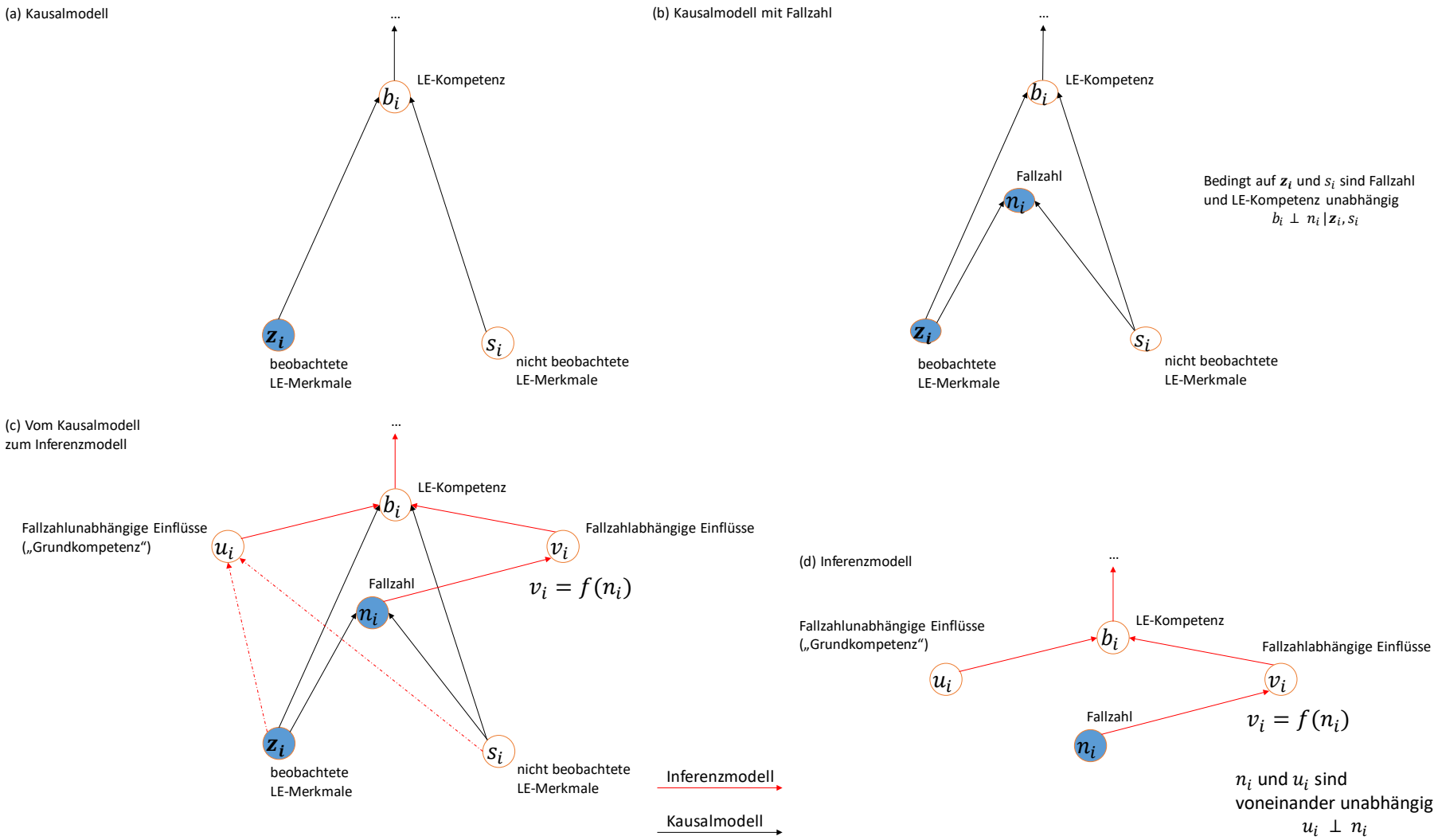

Abbildung 4: (a) bis (d): Ausschnitt der Leistungserbringereinflüsse des grafischen Modells aus Abbildung 3. Übergang vom Kausalmodell zum Inferenzmodell

4.1.3 Vollständiges Inferenzmodell

Abbildung 5 zeigt das in den folgenden Analysen verwendete Inferenzmodell zur Inferenz eines potentiellen Fallzahl-Ergebnis-Zusammenhangs. Anders als in Abbildung 3 enthält das Modell noch eine dritte Ebene: Neben der Patientenebene (grüne Box) und Leistungserbringer-Ebene (orange Box) gibt es eine weitere Populationsebene (gelbe Box) für Variablen, die als leistungserbringerübergreifend konstant modelliert werden. Die Leistungserbringerebene entspricht dabei den Knoten aus Abbildung 4(d). Anders als das Kausalmodell beschreibt das Inferenzmodell nur statistische Zusammenhänge.[2] Das Patienten-Outcome y_{ij} wird als eine Zufallsvariable

$$y_{ij} \sim \mathrm{Ber}(\pi_{ij})$$

mit einer Ereigniswahrscheinlichkeit π_{ij} modelliert, die sich wiederum aus einem Patientenanteil und einem Leistungserbringeranteil zusammensetzt. Dabei werden die Ergebnisse y_{ij} bedingt auf die Einflussgrößen η_{ij}, n_i, u_i nach Gleichung (1) als unabhängige Zufallsvariablen modelliert.[3] Durch diese Betrachtung von y_{ij} als Zufallsvariable können die sonstigen Einflüsse c_{ij} als zufällige Störeinflüsse modelliert werden (vgl. IQTIG 2019b: 180 ff.), treten aber nicht mehr als eigener Modellparameter im grafischen Modell auf.

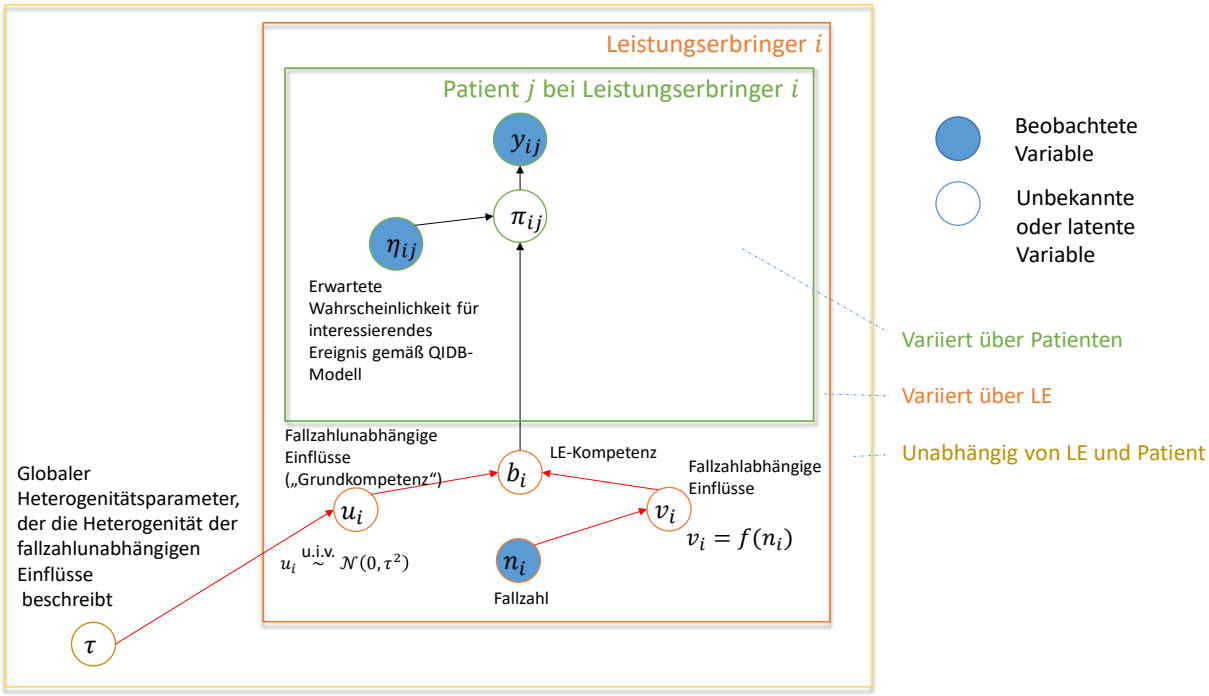

Abbildung 5: Grafisches Modell zur Inferenz eines potentiellen Zusammenhangs von Fallzahl und Patienten-Outcome

[2] Die Richtung der Pfeile wird dennoch beibehalten, stellt hier allerdings keine Kausalzusammenhänge dar, sondern den für die Berechnung nach Formel (1) erforderlichen Informationsfluss.
[3] Insbesondere die Modellierung der Wahrscheinlichkeit bedingt auf die Random-Intercepts u_i ist dabei hervorzuheben. Es wird im Sinne von Fahrmeir et al. (2009) (S. 280 f.) nicht der marginale sondern der konditionale Erwartungswert der Zufallsvariablen y_{ij} modelliert.

Ein weiterer Unterschied zum Kausalmodell besteht darin, dass der Zusammenhang zwischen Patienten-Outcome und Patientenmerkmalen nicht mehr explizit modelliert wird, sondern als gegebene Größe η_{ij} aus dem QIDB-Risikoadjustierungsmodell (vgl. AQUA 2015b) als bekannt vorausgesetzt wird, denn es ist nicht das Ziel des Inferenzmodells, den Zusammenhang von Patientenmerkmalen und Patienten-Outcome neu zu untersuchen. Mit η_{ij} bezeichnen wir dabei konkret die Logit-transformierten patientenspezifischen Risikoscores, die sich aus dem jeweiligen QI-spezifischen QIDB-Risikoadjustierungsmodell (vgl. AQUA 2015b) ergeben:[4]

$$\eta_{ij} = \sum_{m=1}^{M'} \hat{\beta}^m x_{ij}^m$$

Hierbei sind $\hat{\beta}^m, m = 1, \ldots, M'$ die von AQUA (2015b) geschätzten Regressionskoeffizienten für die in der Risikoadjustierung berücksichtigten patientenseitigen Risikofaktoren x_{ij}^m.

Auf Leistungserbringerebene finden sich die in Abschnitt 4.1.2 eingeführten Knoten u_i und v_i, die die Kompetenz des Leistungserbringers in einen fallzahlabhängigen und einen fallzahlunabhängigen Teil aufteilen. Schließlich wurde auf Populationsebene ein weiterer Knoten τ eingeführt. Dieser Knoten wurde eingeführt, weil die Grundkompetenz u_i als sogenannter Random-Effect (vgl. u. a. Agresti 2013: S. 489 ff., Stroup 2016: S. 38 ff., Gelman und Hill 2007: S. 245 ff.) modelliert wird, mit einer Normalverteilungsannahme:

$$u_i \overset{\text{u.i.v.}}{\sim} \mathcal{N}(0, \tau^2)$$

Eine solche Verteilungsannahme ist notwendig, da in Fixed-Effects-Modellen, in denen alle Effekte, insbesondere u_i, als feste Größen und nicht als Zufallsvariablen in das Modell eingehen, aufgrund von Identifizierbarkeitsproblemen kein Fallzahleffekt geschätzt werden kann. Grundsätzlich lassen sich in Fixed-Effects-Modellen keine Effekte von Einflussgrößen ermitteln, für die es für einen Leistungserbringer nur einen Wert gibt, wie es bei der Fallzahl der Fall ist (vgl. u. a. Townsend et al. 2013). Darüber hinaus hat die Verteilungsannahme den Vorteil, dass die Schätzung der Parameter u_i stabilisiert wird und auf diese Weise numerische Probleme vermeidet, die sich ergeben würden, wenn ein Leistungserbringer keine Patientinnen und Patienten behandelt hat, bei denen ein interessierendes Ereignis (z. B. eine Komplikation) aufgetreten ist, was in den betrachteten Daten eine häufige Konstellation ist. Des Weiteren bietet der Parameter τ^2 als *Between-Provider-Variance* des fallzahlunabhängigen Anteils an der Leistungserbringerkompetenz ein Maß für die Heterogenität der Leistungserbringer abzüglich der fallzahlabhängigen Komponente. Somit kann der geschätzte Parameter $\hat{\tau}^2$ verwendet werden, um die Relevanz des inferierten Fallzahl-Ergebnis-Zusammenhangs f mit der Heterogenität in der fallzahlunabhängigen Grundkompetenz der Leistungserbringer ins Verhältnis zu setzen. Zur Interpretation des Heterogenitätsparameters τ^2 bieten sich dabei sog. Median-Odds-Ratios an (vgl. Larsen et al. 2000). Würde man zwei Patientinnen oder Patienten mit identischen Risikoscores von zwei zufällig ausgewählten Leistungserbringern A und B mit gleicher Fallzahl behandeln lassen, so wäre

[4] Im Allgemeinen ist davon auszugehen, dass die in der Risikoadjustierung berücksichtigten Risikofaktoren nicht deckungsgleich mit den theoretisch relevanten Einflussgrößen des Kausalmodells sind, da unter anderem aus Gründen der Datensparsamkeit nicht alle Risikofaktoren erhoben worden sind.

der Odds-Ratio zwischen der Behandlung bei Leistungserbringer A mit dem höheren Risiko und der Behandlung bei Leistungserbringer B mit dem geringeren Risiko im Median:

$$\mathrm{MOR}(\tau) = \exp\left(\sqrt{2 \cdot \tau^2} \cdot \Phi^{-1}(3/4)\right) \tag{2}$$

Dabei ist Φ^{-1} die inverse Verteilungsfunktion der Standardnormalverteilung. Der MOR ist per Konstruktion immer größer als 1, wird aber für die Ergebnisdarstellung in Kapitel 5 bei monoton fallenden Fallzahl-Ergebnis-Verläufen invertiert, um auch in diesem Fall eine hilfreiche Vergleichsgröße zu bieten. Setzt man anstelle der unbekannten Varianz τ^2 die Schätzung $\hat{\tau}^2$ ein, kann der Median-Odds-Ratio dann mit geschätzten Odds-Ratios anderer Risikofaktoren (beispielsweise aus der Risikoadjustierung) verglichen werden, mit dem Ziel beispielsweise fallzahlabhängige und fallzahlunabhängige Einflussgrößen miteinander vergleichen zu können.

Die Modellgleichung für das vollständige Inferenzmodell, hergeleitet aus dem grafischen Modell in Abbildung 5, lautet wie in Gleichung (1):

$$\mathrm{Logit}\big(\mathrm{P}\big(y_{ij} = 1 \big| \eta_{ij}, n_i, u_i\big)\big) = \beta_0 + \eta_{ij} + f(n_i) + u_i,$$

mit Random-Effects $u_i \overset{\text{u.i.v.}}{\sim} \mathcal{N}(0, \tau^2)$, die wie im grafischen Modell in Abbildung 5 als unabhängig von der Fallzahl angenommen werden. Der Parameter β_0 ist ein globaler Intercept-Parameter, der bewirkt, dass die Random-Effects u_i um Null zentriert sind. Ziel des Inferenzmodells ist es, Kenntnis über den Einfluss der Fallzahl $f(n_i)$ auf die fallübergreifende Leistungserbringerkompetenz und somit auf die Wahrscheinlichkeit für das Eintreten des interessierenden Ereignisses zu erlangen, während für die restlichen Modellkomponenten (insbesondere die patientenseitigen Risikofaktoren) kontrolliert wird. Die Risikoscores η_{ij} werden dabei nicht vom Modell neu geschätzt, sondern wurden von AQUA (2015b) übernommen und gehen als sogenannte Offset-Parameter in das Modell ein.

Nicht im Modell berücksichtigt werden Interaktionen von η_{ij} mit n_i oder u_i, die eine besondere Leistungserbringerkompetenz für bestimmte Patientengruppen modellieren könnten. Diese Interaktionen werden aus Gründen der zu hohen Modellkomplexität vernachlässigt, da die Seltenheit der betrachteten Outcomes in der Datengrundlage keine verlässliche Schätzung solcher Interaktionen zuließe. Das Nichtvorhandensein von Interaktionstermen sollte jedoch bei der Interpretation der Ergebnisse bedacht werden (vgl. Kapitel 6). Es wäre theoretisch denkbar, dass sich der Fallzahl-Ergebnis-Zusammenhang bei Patienteninnen und Patienten mit hohem Risiko bzgl. des Outcomes anders darstellt als bei Patientinnen und Patienten mit niedrigem Risiko.

4.2 Schätzung des Fallzahleffekts f

Das zentrale interessierende Objekt ist die Funktion f für jeden der acht Qualitätsindikatoren.

Um eine Schätzung $\hat{f}$ zu erhalten bedienen wir uns eines flexiblen Glättungsansatzes durch Splines: Dazu wird der Bereich der Fallzahlen in eine gewisse Anzahl von möglicherweise verschieden großen Teilstücken zerlegt (gesteuert durch die Anzahl und Position der sogenannten Knoten); auf diesen Teilstücken approximiert man die unbekannte Funktion f derart durch Polynome, dass die resultierende zusammengesetzte Funktion $\hat{f}$ glatt ist und keine sehr erratische Form hat, siehe Fahrmeir et al. (2009) (dort Kapitel 7) für eine Einführung.

Der zentrale Vorteil dieser glatten Modellierung besteht darin, dass eine große Bandbreite von Zusammenhängen abgedeckt ist: Die klassische lineare Modellierung stellt eine sehr starke, im Allgemeinen nicht realistische Einschränkung dar – aus der Linearität (auf der Ebene der rechten Seite von (1)) folgt zum Beispiel zwangsläufig die Monotonie der geschätzten Zusammenhänge. Ein Spline-Ansatz kann hingegen auch nicht monotone Zusammenhänge abbilden, ohne dabei die Möglichkeit einfacher, linearer Verläufe auszuschließen. Eine Analyse durch Bildung von Volumengruppen lässt sich als Spezialfall des Spline-Ansatzes betrachten: Auf den oben genannten Teilstücken wird f durch Konstanten, das heißt Polynome 0'ten Grades, geschätzt, wobei man aber auf die Glattheit der Gesamtkurve verzichtet. Die dadurch entstehenden abrupten Sprünge sind im Allgemeinen ebenfalls nicht realistisch und der Wegfall der Glattheitsbedingung verursacht eine besonders starke Empfindlichkeit der Schätzung bezüglich der Knotenwahl. Der Ausschluss plötzlicher Sprünge vermeidet außerdem eine zu starke Anpassung an die Daten (Overfitting).

Grundsätzlich wird f gemeinsam mit den anderen unbekannten Modellparametern durch einen penalisierten Maximum-Likelihood-Ansatz geschätzt. Eine genaue Beschreibung der Verfahren findet sich in (Wood 2017).

Die Bandbreite der Möglichkeiten hinsichtlich der Spline-Basis, Knotenzahl und -position, des Grades sowie der Form des Bestrafungsterms, ist groß. Das IQTIG hat verschiedene Varianten verglichen (siehe Abschnitt 5.4) und für die im nächsten Abschnitt präsentierten Ergebnisse folgende Spezifikation gewählt: Der Fallzahlbereich wird in sieben Teilstücke zerlegt, die gleich viele der vorkommenden Fallzahlen enthalten (quantilbasierte Knotensetzung). Die verwendeten Polynome dritten Grades basieren auf einer „cardinal"-Spline-Basis. Eine starke Krümmung der geschätzten Funktion $\hat{f}$ auf dem Bereich $[n_{\min}, n_{\max}]$ der Fallzahlen wird durch den Bestrafungsterm

$$\lambda \int_{n_{\min}}^{n_{\max}} f''(x)^2 \, \mathrm{d}x$$

kontrolliert, wobei der sogenannte Glättungsparameter λ den Einfluss des Bestrafungsterms und somit die Glattheit der geschätzten Kurve kontrolliert. Der Glättungsparameter wird so aus den Daten geschätzt, dass die geschätzte Kurve die Daten möglichst gut beschreibt, gleichzeitig aber keine Überanpassung (Overfitting) erfolgt. Für technische Details siehe Wood (2017) (dort Abschnitt 5.3.1).

4.3 Modellwahlkriterien und statistische Bewertung der Ergebnisse

Vor der statistischen Bewertung eines möglichen Fallzahleffekts sollte zunächst ermittelt werden, ob sich das gewählte Modell mit den gewählten Parametern für die Untersuchung eignet. Zum einen ziehen wir dazu die klassische sogenannte AUC (area under the curve) heran: ein Maß zwischen 0,5 und 1, welches umso näher an 1 liegt, je treffsicherer das geschätzte Modell die interessierenden Ereignisse vorhersagen kann, siehe etwa Hosmer et. al (2013) (dort Abschnitt 5.2.4). Zum anderen betrachten wir das konditionale AIC (Akaike Information Criterion) wie in

Wood et al. (2016) vorgestellt: ein positives relatives Maß, das die Modellanpassung gegen die Modellkomplexität abwägt.

Wir möchten für jeden Qualitätsindikator eine statistische Bewertung dafür erhalten, ob es einen nichttrivialen Einfluss der Fallzahl auf π_{ij} gibt. Dies entspricht in unserem Rahmen dem Testproblem

$$H_0\colon\ f \equiv 0 \ \text{ gegen } \ H_1\colon\ f \not\equiv 0,$$

also der Nullhypothese, dass f konstant gleich 0 ist, gegen die Alternativhypothese, dass f irgendwo im Fallzahlbereich von 0 abweicht. Wie in der Literatur (Wood 2012, Marra und Wood 2012) beschrieben, lässt sich ein geeigneter statistischer Test zugehörig zu bestimmten sogenannten simultanen Konfidenzbändern um die geschätzte Funktion $\hat{f}$ formulieren. Dabei wird die Unsicherheit in der Schätzung von β_0 und der Random-Effects mitbedacht. Zur Entscheidung über statistische Signifikanz verwenden wir das Niveau $\alpha = 5\%$.

Es ist zu beachten, dass ein in diesem Sinne statistisch signifikantes Ergebnis keinen besonders starken Fallzahl-Ergebnis-Zusammenhang signalisiert, sondern dass die Daten unter den Modellannahmen in der Hypothese widersprechen, dass es *keinen* Zusammenhang gibt. Dies wird auch angesichts der sehr kleinen in Tabelle 2 aufbereiteten Prävalenzen deutlich: Große absolute Unterschiede sind ohnehin – selbst bei einem signifikanten Zusammenhang – ausgeschlossen. Schon allein aus diesem Grund muss eine abschließende Bewertung der Ergebnisse neben der statistischen Signifikanz auch die klinische Relevanz einbeziehen.

4.4 Abgrenzung zu anderen Methoden der Volumen-Outcome-Analyse

In der Literatur finden sich unterschiedliche methodische Ansätze für die Analyse von Volumen-Outcome-Zusammenhängen, die sich insbesondere hinsichtlich der Berücksichtigung der Clusterstruktur aufgrund der stochastischen Abhängigkeit von Outcomes, die zum gleichen Standort gehören, sowie der Modellierung der Fallzahl unterscheiden (Wetzel 2006).

Wird die Clusterstruktur nicht berücksichtigt, kann dies insbesondere dazu führen, dass die Konfidenzintervalle zu schmal geschätzt werden und Ergebnisse fälschlicherweise als signifikant eingestuft werden (vgl. Urbach und Austin (2005), Wetzel (2006) und Fahrmeir et al. (2009), dort Kapitel 6). Es gibt unterschiedliche Möglichkeiten Clusterstrukturen in Daten zu berücksichtigen. Für die vorliegenden Analysen wurden sogenannte konditionale hierarchische Modelle gewählt, die eine weit verbreitete Modellklasse im Umgang mit Clusterdaten darstellen (siehe z. B. Fahrmeir et al. (2009) für eine Einführung). Der Random-Intercept- bzw. Grundkompetenz-Parameter u_i in Modellgleichung (1) stellt dabei formal eine Möglichkeit dar, die stochastische Abhängigkeit von Outcomes zu modellieren, die zum gleichen Standort gehören.

Ebenfalls konditionale hierarchische Modelle zur Analyse von Volumen-Outcome-Zusammenhängen werden beispielsweise in George et al. (2017) verwendet. Insbesondere wird ein vergleichbarer penalisierter Spline-Ansatz mit vorheriger Log-Transformation der Fallzahl vorgestellt. George et al. (2017) erweitern dabei den Modell-Ansatz auch um eine möglicherweise existierende fallzahlabhängige Heterogenität der Grundkompetenz. Während George et al.

(2017) einen vollständig Bayesianischen Ansatz und daher die Bestimmung aller Modellkoeffizienten über einen Markov-Chain-Monte-Carlo-Ansatz durchführen, wird im vorliegenden Bericht ein (weitestgehend) frequentistischer Ansatz mit penalisierter Maximum-Likelihood-Optimierung verwendet.

Varagunam et al. (2015) verwenden ebenfalls ein frequentistisches konditionales hierarchisches Modell mit Random-Intercepts für die Leistungserbringer. Im Gegensatz zur in diesem Bericht verwendeten Analyse bilden sie jedoch Volumengruppen anstelle den Einfluss der stetigen Fallzahl zu modellieren.

Grundsätzlich existiert noch (mindestens) eine weitere Möglichkeit, die Abhängigkeitsstruktur der Outcomes in der Modellierung zu berücksichtigen. Es handelt sich dabei um sogenannte marginale Modelle (in Abgrenzung zu den hier gewählten konditionalen Modellen), bei denen anstelle des zufälligen Intercepts eine Annahme über die Kovarianzstruktur der Daten, die von einem Standort stammen, getroffen wird, siehe Fitzmaurice et al. (2011) (Kapitel 7) für eine Einführung. Beispiele für die Verwendung marginaler Modelle zur Analyse von Volumen-Outcome-Zusammenhängen finden sich in IQWIG (2006).

Das IQTIG hat auch die Verwendung marginaler Modelle ausführlich geprüft, jedoch letztlich aus folgenden Gründen das konditionale Modell, also das oben erläuterte Modell, ausgewählt: Aus der Literatur geht keine klare Empfehlung bezüglich der Verwendung von marginalen gegenüber konditionalen Modellen für derartige Analysen hervor. Jedoch weisen aktuelle Artikel darauf hin, dass die bisher in der Literatur als Vorteil genannte einfachere Interpretation von marginalen Modellen nicht immer korrekt ist, siehe Muff et al. (2016) und auch Lee und Nelder (2004). Darüber hinaus lässt sich das konditionale Modell, wie in Abschnitt 4.1 beschrieben, direkt aus einem grafischen Modell herleiten, in dem die getroffenen Annahmen transparent gemacht werden. Insbesondere in Kombination mit der flexiblen Modellierung der Fallzahl mithilfe von penalisierten Splines bietet die Verwendung von konditionalen Modellen etablierte Methoden und robuste, flexible Software für die Schätzung und zur Beurteilung der statistischen Signifikanz. Speziell verwendet das IQTIG die Sprache R (vgl. R Core Team (2019)) und die Modelle sind im R-Paket `mgcv` implementiert (vgl. Wood 2019).

Hinsichtlich der Modellierung der Fallzahl kann insbesondere danach unterschieden werden, ob die kontinuierliche Fallzahl direkt verwendet wird (z. B. IQWIG (2006), Grouven et al. (2008), IQWIG (2019)) oder zuvor diskretisiert wird, d. h. die Fallzahl in Form von Volumengruppen in die Analysen eingeht (z. B. Birkmeyer (2002), Stefoski Mikeljevic et al. (2003), Varagunam et al. (2015)).

Das IQTIG hat auch die im Auftrag vorgeschlagene Methode der Bildung von Volumengruppen, also die diskrete Modellierung eines potentiellen Fallzahl-Effektes geprüft. Dabei gibt es mehrere Quellen (vgl. u. a. Grouven et al. 2008, Wetzel 2006, IQWIG 2019), die von einer Einteilung der Leistungserbringer in feste Volumengruppen abraten. Die Gründe, die dabei genannt werden, sind folgende: Die Wahl spezifischer Volumengruppen bedarf immer einer Setzung. Die Ergebnisse können mitunter erheblich von den gewählten Volumengrenzen abhängen. Des Weiteren wird bei der Diskretisierung der stetigen Fallzahl in Fallzahlgruppen implizit die Annahme

getroffen, dass sich der zu messende Fallzahleffekt innerhalb dieser Gruppen nicht unterscheidet oder einen vorgegebenen Verlauf hat, was mit einem Informationsverlust innerhalb der Gruppen einhergeht. An den Volumengrenzen wiederum entstehen unplausible Brüche: Der Effekt von Fallzahlen geringfügig oberhalb einer Volumengrenze auf das Outcome wird möglicherweise drastisch anders geschätzt als der von Fallzahlen geringfügig unterhalb der gleichen Volumengrenze. Auch die in Abschnitt 4.2 beschriebenen Glättungsmethoden kommen nicht ohne Annahmen und Spezifikation weiterer Modellparameter aus. Beispielsweise muss eine konkrete Spline-Basis oder die Anzahl der Knotenpunkte gewählt werden. Durch die Bestrafung komplexer Kurvenverläufe erweisen sich die Glättungsmethoden allerdings als weniger sensitiv gegenüber diesen genauen Parameterspezifikationen als rein deskriptive Auswertungen in Fallzahlgruppen (vgl. Abschnitt 5.4). Nach Abwägung dieser Vor- und Nachteile zwischen stetiger und diskreter Modellierung des Fallzahleffektes hat sich das IQTIG für den penalisierten Spline-Ansatz entschieden. Als Sensitivitätsanalysen wurden auch Modellrechnungen mit verschiedenen Volumengruppen durchgeführt (vgl. Abschnitt 5.4).

5 Ergebnisse

Im Folgenden werden Ergebnisse einzeln für jeden Qualitätsindikator in Form von numerischen Ergebnissen und Grafiken dargestellt und interpretiert. Zunächst enthält jeweils eine Tabelle allgemeine Informationen zum Qualitätsindikator sowie einige ermittelte statistische Kennzahlen: den *p*-Wert zum Test, ob der Fallzahleffekt statistisch signifikant ist (siehe Abschnitt 4.3), den Median-Odds-Ratio (siehe Gleichung (2) in Abschnitt 4.1.3) zur Einschätzung der Variabilität zwischen den Standorten sowie die AUC (siehe Abschnitt 4.3). Zudem wird jeweils eine Grafik mit logarithmischer Skala und einer Tabelle mit Odds-Ratios für eine Auswahl an Fallzahlpaaren angegeben. Für den ersten Qualitätsindikator soll dies exemplarisch vorgestellt werden.

5.1 Qualitätsindikatoren, bei denen statistische Signifikanz vorliegt

5.1.1 QI 51054: Wundhämatome/Nachblutungen

Tabelle 3: Zusammenfassung QI 512054 (Wundhämatome)

Zähler	Operationen, bei denen beim Patienten ein Wundhämatom oder eine Nachblutung auftrat.
Nenner	Alle Operationen bei Patienten ab 20 Jahre.
Risikoadjustierung	Geschlecht, ASA-Klassifikation.
Statistische Ergebnisse	p-Wert: 0,033, $\widehat{\mathrm{MOR}}$: 0,529, In-Sample-AUC: 0,787.

Da der *p*-Wert hier die Schwelle $0{,}05$ unterschreitet, wird der Fallzahleffekt als statistisch signifikant eingestuft. Der bereits erläuterte MOR liegt deutlich unter 1, was eine relativ große Streuung der Standort-Grundkompetenzen signalisiert. Schließlich ist die AUC groß genug, um von einer ausreichenden Prädiktionskraft des Modells auszugehen – als Richtwert für diese Einschätzung wird die untere Schwelle 0,7 genannt, siehe (Hosmer et. al 2013, Abschnitt 5.2.4).

Im Folgenden soll der Zusammenhang zwischen $\hat{f}$ und der resultierenden Wahrscheinlichkeit des interessierenden Ereignisses (hier also das Vorkommen von Wundhämatomen oder Nachblutungen) grafisch dargestellt werden. Da diese Wahrscheinlichkeit jedoch stets von der Standort-Grundkompetenz u_i und gegebenenfalls vom individuellen Risikoscore η_{ij} abhängt, müssen diese Werte für eine explizite Darstellung fest gewählt werden. Es wird dazu einen Standort mit durchschnittlicher Grundkompetenz $u^* = 0$ und ein Fall mit Median-Risikoscore $\eta^* = \mathrm{median}(\eta_{ij})$ gewählt. Dies hat unter Umständen einen geringfügigen Einfluss auf die Form und Höhe der Kurve in der Grafik, jedoch nicht auf ihren qualitativen Verlauf (Monotonieverhalten, Größenordnung der angenommenen Werte) und insbesondere auch nicht auf die statistische Inferenz, da die Werte dort nicht eingehen. In die Grafik (hier Abbildung 6: Verlaufsgrafik QI 51054 (Wundhämatome)) werden dann als glatte Kurve (schwarz) die resultierenden Wahrscheinlichkeiten eingetragen:

$$P(y=1|\eta^*,n,u^*)=\begin{cases}\text{Logit}^{-1}\left(\widehat{\beta_0}+\eta^*+\hat{f}(n)+u^*\right), & \text{bei Risikoadjustierung,}\\ \text{Logit}^{-1}\left(\widehat{\beta_0}+\hat{f}(n)+u^*\right), & \text{sonst.}\end{cases}$$

Für diesen ersten Qualitätsindikator trifft also die obere Variante zu.

Zur Einordnung dieses Ergebnisses kommen vier Aspekte hinzu:

1. Ein 95% −Konfidenzband (grün, leicht transparent) für die Kurve wie in Abschnitt 4.3 beschrieben, aber ebenfalls durch Logit^{-1} transformiert.

2. Eine waagerechte Referenzlinie (schwarz), die dem Fall entspricht, dass kein Fallzahl-Ergebnis-Zusammenhang besteht. Sie ist gegeben durch die Konstante

$$P_0(y=1|\eta^*,u^*)=\begin{cases}\text{Logit}^{-1}\left(\widehat{\beta_0}+\eta^*+u^*\right), & \text{bei Risikoadjustierung,}\\ \text{Logit}^{-1}\left(\widehat{\beta_0}+u^*\right), & \text{sonst.}\end{cases}$$

3. Schließlich tragen wir in die Grafik auch die einzelnen erwarteten Wahrscheinlichkeiten je Standort für den gemeinsamen durchschnittlichen Fall ein (blau), also die Punkte

$$\left(n_i\ ,\ P(y_{ij}=1|\eta^*,n_i,\widehat{u_i})\right).$$

   Dies ermöglicht einen klaren Eindruck davon, welche Variabilität in der resultierenden Wahrscheinlichkeit durch die unterschiedlichen Grundkompetenzen verursacht wird.

4. Am oberen Rand ist schließlich ein Histogramm der beobachteten Fallzahlen angegeben.

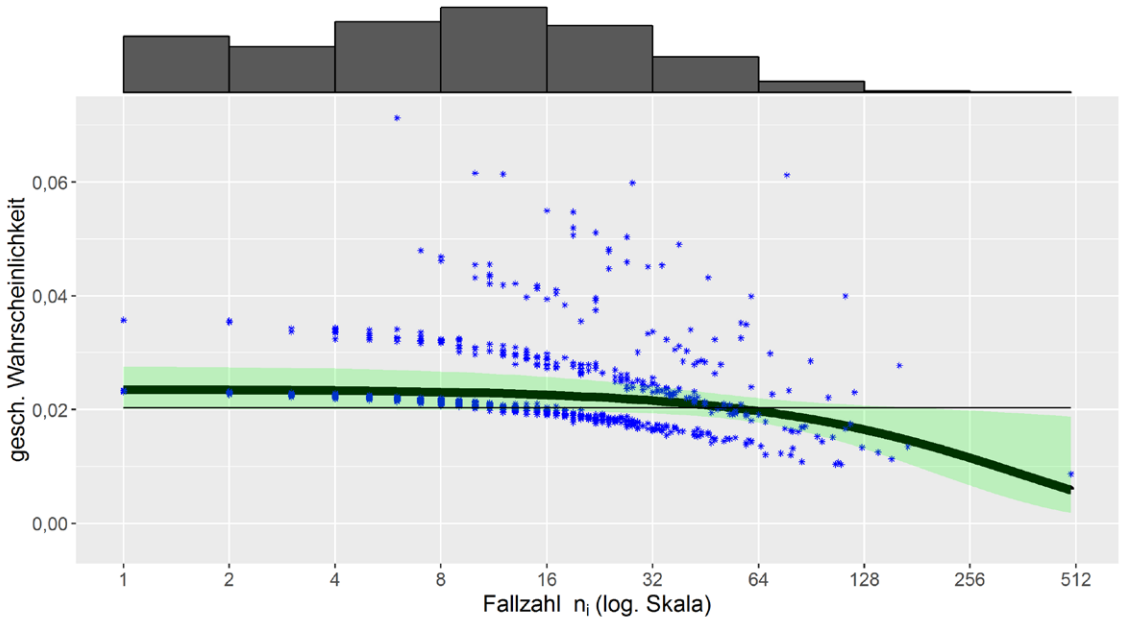

Abbildung 6: Verlaufsgrafik QI 51054 (Wundhämatome)

In diesem Fall ist die geschätzte, bezüglich Grundkompetenz und Risikoscore standardisierte Wahrscheinlichkeit für das Auftreten von Wundhämatomen/Nachblutungen mit der Fallzahl monoton fallend von ca. 2,3 % für die kleinste vorkommende Fallzahl auf 0,5 % für die größte

vorkommende Fallzahl, also in einem sehr kleinen Bereich von Wahrscheinlichkeiten. Bezogen auf diesem Bereich streuen die Standortergebnisse allerdings sehr stark, bis zum dreifachen maximalen Wert der eigentlichen Kurve. Die Hypothese, dass es keinen Fallzahl-Ergebnis-Zusammenhang gibt (d. h. f der waagerechten Linie entspricht), wird abgelehnt, da das Konfidenzband die waagerechte Linie nicht vollständig überdeckt.

An dieser Stelle wird betont, dass es sich nicht um ein punktweises, sondern ein simultanes Konfidenzband handelt, siehe etwa Fahrmeir et al. (2009, Abschnitt 7.1.8), was sich auch auf den Test überträgt: Die Signifikanzeinstufung bezieht sich auf die Kurve im Gesamtverlauf, sodass nicht anhand des Konfidenzbandes ein Teilbereich mit signifikantem Effekt abgelesen werden sollte. Die Tatsache, dass die geschätzte Kurve monoton fällt, bleibt unabhängig davon ein klares Ergebnis der Analyse.

Odds-Ratios-Tabelle

Es wird nun eine weitere Möglichkeit besprochen, die ermittelten Ergebnisse darzustellen. Dafür müssen zwar keine Werte für die Standort-Grundkompetenz und die patientenspezifischen Risikoscores explizit gewählt werden, aber für die Darstellung eignet sich eher eine Tabelle als eine Grafik. Eine besondere Eigenschaft der gewählten Logit-Modellierung ist, dass anhand der geschätzten Ergebnisse leicht Odds-Ratios zwischen verschiedenen Fällen ermittelt werden können. Speziell gilt hier:

Zwischen den Odds des interessierenden Ereignisses bei einem Standort mit Volume n_1 gegenüber einem Standort mit Volume n_2 mit gemeinsamer Grundkompetenz u besteht bei bezüglich Risiko-Score η identischen Fällen in unserem Modell das Verhältnis

$$\frac{\text{Odds}(y = 1|\eta, u, n_1)}{\text{Odds}(y = 1|\eta, u, n_2)} = \exp(f(n_1) - f(n_2)).$$

Die folgenden Ergebnisse sind dementsprechend die geschätzten Odds-Ratios für zwei Fälle mit beliebigem, aber identischem Risikoscore in Standorten mit beliebiger, aber identischer Grundkompetenz und beschreiben dementsprechend den Einfluss der Fallzahl, wenn alle anderen Einflussfaktoren fest sind.

Die für die Tabelle, z. B. Tabelle 4: Odds-Ratios QI 51054 (Wundhämatome), gewählten Fallzahlen entsprechen intuitiv vorstellbaren Häufigkeiten wie „durchschnittlich ein Eingriff pro Quartal“, „durchschnittlich ein Eingriff pro Monat” und so weiter. Man beachte, dass es sich auch hier um geschätzte Werte handelt und auf eine zusätzliche Darstellung der statistischen Unsicherheit verzichtet wird.

Tabelle 4: Odds-Ratios QI 51054 (Wundhämatome)

	Fallzahl n_1				
Fallzahl n_2	**4**	**12**	**24**	**52**	**104**
4	1	0,977	0,944	0,872	0,751
12		1	0,966	0,892	0,768

	Fallzahl n_1				
Fallzahl n_2	**4**	**12**	**24**	**52**	**104**
24			1	0,923	0,795
52				1	0,862
104					1

Als Beispiel lässt sich aus der Tabelle ablesen, dass die Odds für das Auftreten von Wundhämatomen/Nachblutungen bei einem Patienten oder einer Patientin um den Faktor 0,768 besser sind, wenn er/sie in einem Standort mit 104 Fällen gegenüber einem Standort mit nur 12 Fällen, aber der gleichen Grundkompetenz, behandelt wird.

Während man für diese Odds-Ratios also die Fallzahl variieren lässt, während Risikoscore und Grundkompetenz fest sind, geht man beim Median-Odds-Ratio von einem festen Risikoscore und einer festen Fallzahl aus und misst den dann noch zu erwartenden Einfluss der variierenden Grundkompetenz.

Man beachte, dass die Grafik und Tabelle beziehungsweise der MOR tatsächlich unterschiedliche, sich ergänzende Informationen enthalten: In der Grafik sind spezielle Wahrscheinlichkeiten zu jeder einzelnen Fallzahl eingezeichnet, während in der Tabelle jeweils zwei Fallzahlen herangezogen und die beiden resultierenden Odds (nicht Wahrscheinlichkeiten) dividiert werden. Die Tatsache, dass für diesen QI alle Odds-Ratios kleiner als 1 sind, hängt allerdings tatsächlich direkt mit der Monotonie der geschätzten Funktion $\hat{f}$ zusammen.

5.1.2 QI 51064: Reoperationen aufgrund von Komplikationen

Tabelle 5: Zusammenfassung QI 51064 (Reoperationen)

Zähler	Operationen, nach denen der Patient aufgrund von Komplikationen reoperiert werden musste.
Nenner	Alle Operationen bei Patienten ab 20 Jahren.
Risikoadjustierung	Geschlecht, Alter, Wundkontaminationsklassifikation, Entzündungszeichen im Labor, positiver Erregernachweis, Indikation periprothetische Fraktur.
Statistische Ergebnisse	p-Wert: 0,0095, $\widehat{\mathrm{MOR}}$: 0,579, AUC: 0,782.

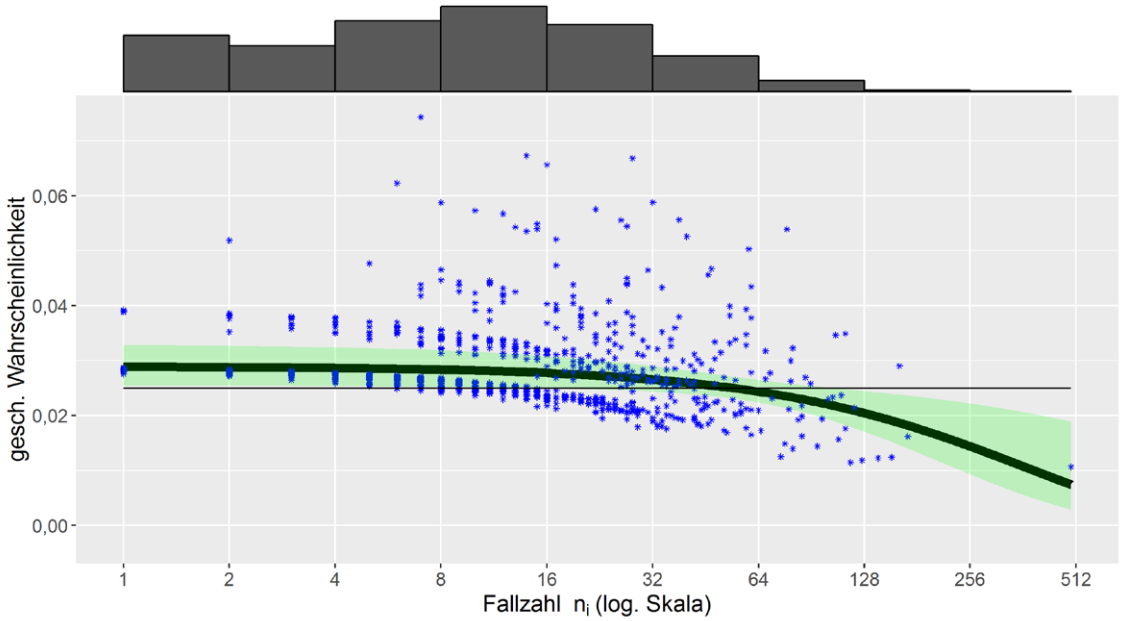

Abbildung 7: Verlaufsgrafik QI 51064 (Reoperationen)

Tabelle 6: Odds-Ratios QI 51064 (Reoperationen)

	Fallzahl n_1				
Fallzahl n_2	**4**	**12**	**24**	**52**	**104**
4	1	0,978	0,946	0,874	0,756
12		1	0,967	0,894	0,773
24			1	0,925	0,799
52				1	0,865
104					1

Die Interpretation der Ergebnisse fällt hier sehr ähnlich aus. Erwähnenswert ist der wesentlich kleinere *p*-Wert (hier wäre der Zusammenhang auch bei Niveau α = 1 % als signifikant eingestuft worden), der sich auch dadurch ausdrückt, dass die waagerechte Linie deutlicher das Konfidenzband verfehlt.

5.1.3 QI 51069: Todesfälle

Tabelle 7: Zusammenfassung QI 51069 (Todesfälle)

Zähler	Verstorbene Patienten.
Nenner	Alle Patienten ab 20 Jahren.

Risikoadjustierung	ASA-Klassifikation, Entzündungszeichen im Labor, Indikation periprothetische Fraktur.
Statistische Ergebnisse	p-Wert: 0,0218, $\widehat{\mathrm{MOR}}$: 0,538, AUC: 0,883.

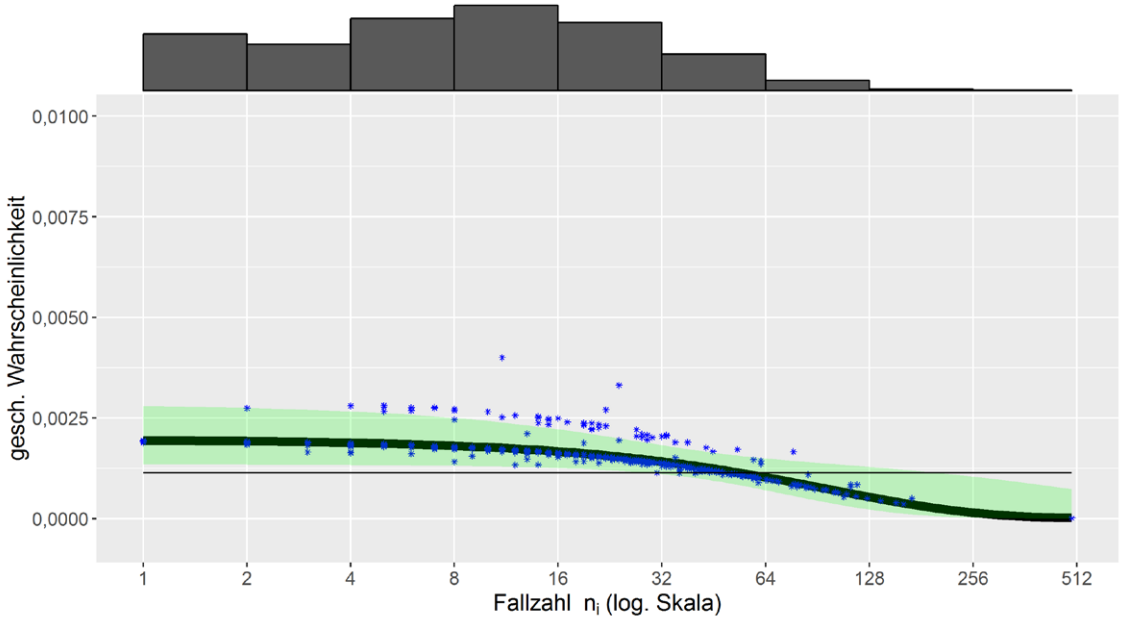

Abbildung 8: Verlaufsgrafik QI 51069 (Todesfälle)

Tabelle 8: Odds-Ratios QI 51069 (Todesfälle)

	Fallzahl n_1				
Fallzahl n_2	**4**	**12**	**24**	**52**	**104**
4	1	0,922	0,817	0,615	0,364
12		1	0,886	0,667	0,394
24			1	0,753	0,445
52				1	0,591
104					1

Hier zeigt sich eine Besonderheit der Odds-Ratios: Die Prävalenzen sind bei diesem QI besonders klein – diese relative Nähe zu 0 sorgt dafür, dass auch sehr kleine Unterschiede zwischen den beiden jeweils eingehenden Wahrscheinlichkeiten zu verhältnismäßig kleinen Odds-Ratio-Werten führen. Dies unterstreicht noch einmal die Tatsache, dass Odds nicht mit Wahrscheinlichkeiten zu verwechseln sind.

Die Interpretation der Ergebnisse für diesen QI ist ansonsten ähnlich zu den vorherigen beiden Qualitätsindikatoren.

5.2 Qualitätsindikatoren, bei denen keine statistische Signifikanz vorliegt

Unter der Annahme, dass f jeweils konstant ist, sind die beobachteten Daten bei den übrigen fünf Qualitätsindikatoren aus statistischer Sicht plausibel, sodass diese Annahme jeweils nicht abgelehnt wird.

5.2.1 QI 51044: Gehunfähigkeit bei Entlassung

Tabelle 9: Zusammenfassung QI 51044 (Gehunfähigkeit)

Zähler	Patienten, die bei der Entlassung nicht selbstständig gehfähig sind und die vor der Operation selbstständig gehfähig waren
Nenner	Alle Patienten ab 20 Jahre, die lebend entlassen wurden
Risikoadjustierung	Alter, Wundkontaminationsklassifikation vor der ersten Operation, Indikation periprothetische Fraktur.
Statistische Ergebnisse	p-Wert: 0,1504, $\widehat{\mathrm{MOR}}$: 0,532, AUC: 0,87.

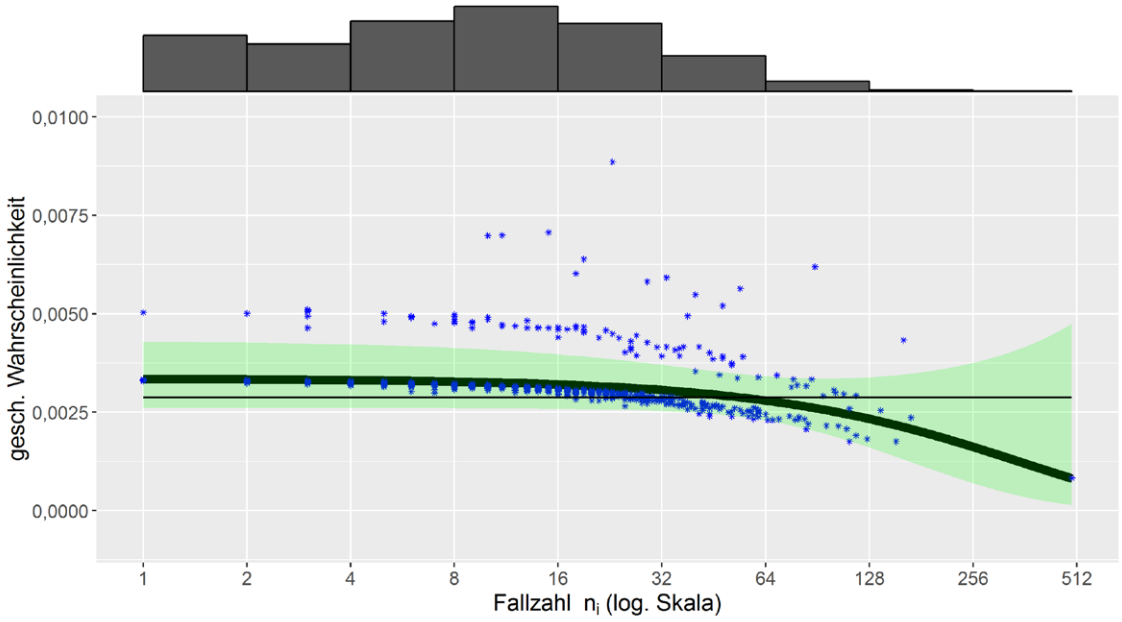

Abbildung 9: Verlaufsgrafik QI 51044 (Gehunfähigkeit)

Der geschätzte Zusammenhang ist auch bei diesem Qualitätsindikator monoton fallend, jedoch wird der Effekt nicht als signifikant eingestuft und die Unsicherheit insbesondere am rechten Rand der Grafik ist sehr groß.

Tabelle 10: Odds-Ratios QI 51044 (Gehunfähigkeit)

	Fallzahl n_1				
Fallzahl n_2	**4**	**12**	**24**	**52**	**104**
4	1	0,977	0,945	0,872	0,752
12		1	0,966	0,892	0,769
24			1	0,923	0,796
52				1	0,862
104					1

5.2.2 QI 51049: Frakturen

Tabelle 11: Zusammenfassung QI 51049 (Frakturen)

Zähler	Operationen, bei denen beim Patienten eine Fraktur auftrat.
Nenner	Alle Operationen bei Patienten ab 20 Jahren.
Risikoadjustierung	Indikation periprothetische Fraktur.
Statistische Ergebnisse	p-Wert: 0,7061, $\widehat{\mathrm{MOR}}$: 0,457, AUC: 0,926.

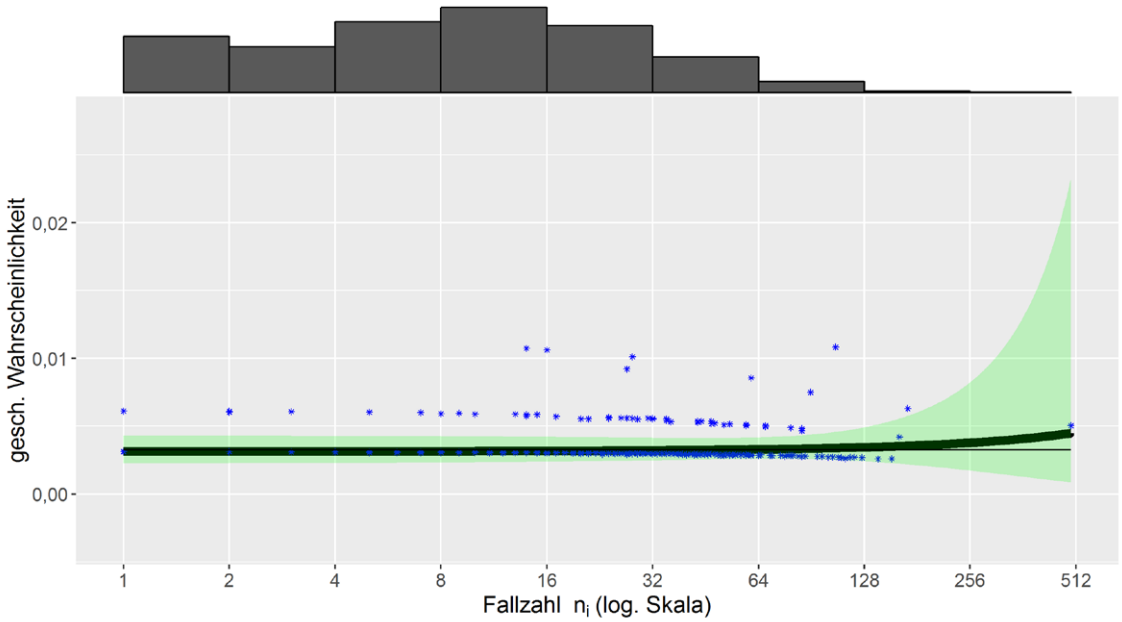

Abbildung 10: Verlaufsgrafik QI 51049 (Frakturen)

Es handelt sich hier um einen von zwei Qualitätsindikatoren, bei denen ein monoton steigender Zusammenhang geschätzt wird. Dieses Ergebnis lässt aufgrund des hohen *p*-Wertes jedoch aus

statistischer Sicht keine Schlussfolgerung über zum Beispiel schlechtere Qualität bei hohen Fallzahlen zu; die geschätzte Wahrscheinlichkeit erhöht sich ohnehin nur sehr leicht. Wie im vorigen Fall beachte man auch die große Unsicherheit bei den besonders hohen Fallzahlen.

Tabelle 12: Odds-Ratios QI 51049 (Frakturen)

	Fallzahl n_1				
Fallzahl n_2	**4**	**12**	**24**	**52**	**104**
4	1	1,006	1,015	1,035	1,075
12		1	1,009	1,029	1,069
24			1	1,02	1,06
52				1	1,038
104					1

Die Monotonie verursacht hier auch Odds-Ratios, die über 1 liegen, siehe zum Vergleich die Bemerkung nach Tabelle 4: Odds-Ratios QI 51054 (Wundhämatome).

Bei diesem Qualitätsindikator ergibt sich zudem der kleinste $\widehat{\mathrm{MOR}}$, was bedeutet, dass sich hier die größte Variabilität der Grundkompetenzen ergibt. Zur Interpretation ist dabei auch hier zu beachten, dass es sich beim $\widehat{\mathrm{MOR}}$ lediglich um eine Schätzung handelt.

5.2.3 QI 51059: Allgemeine postoperative Komplikationen

Tabelle 13: Zusammenfassung QI 51059 (Komplikationen)

Zähler	Patienten mit Pneumonie, kardiovaskulären Komplikationen, tiefer Bein-/Beckenvenenthrombose oder Lungenembolie
Nenner	Alle Patienten ab 20 Jahren.
Risikoadjustierung	Alter, ASA-Klassifikation, Wundkontaminationsklassifikation, Indikation periprothetische Fraktur.
Statistische Ergebnisse	p-Wert: 0,5229, $\widehat{\mathrm{MOR}}$: 0,595, AUC: 0,812.

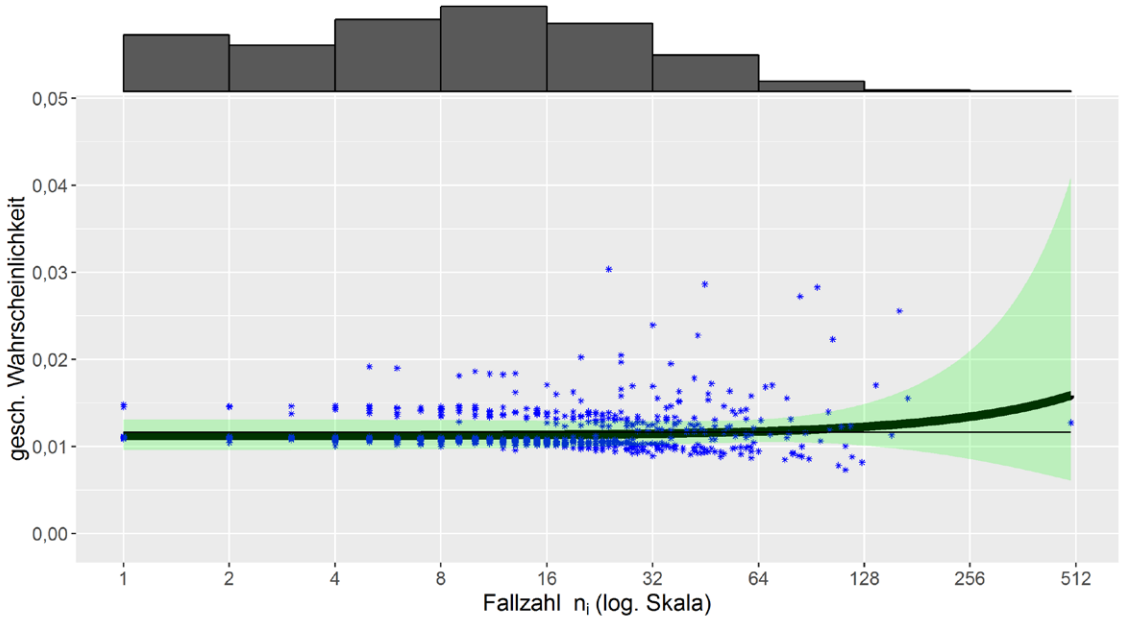

Abbildung 11: Verlaufsgrafik QI 51059 (Komplikationen)

Die Interpretation der Ergebnisse für diesen QI ist analog zum vorigen QI, Abschnitt 5.2.2.

Tabelle 14: Odds-Ratios QI 51059 (Komplikationen)

	Fallzahl n_1				
Fallzahl n_2	**4**	**12**	**24**	**52**	**104**
4	1	1,006	1,015	1,035	1,075
12		1	1,009	1,029	1,068
24			1	1,02	1,059
52				1	1,038
104					1

5.2.4 QI 2220: Gefäßläsionen/Nervenschäden

Abschließend werden die Ergebnisse zu den beiden Qualitätsindikatoren betrachtet, für die kein Risikoscore zur Verfügung steht.

Tabelle 15: Zusammenfassung QI 2220 (Gefäßläsion)

Zähler	Operationen, bei denen beim Patienten eine Gefäßläsion oder ein Nervenschaden auftrat.
Nenner	Alle Operationen bei Patienten ab 20 Jahren.
Risikoadjustierung	Nicht vorhanden.
Statistische Ergebnisse	p-Wert: 0,3607, $\widehat{\mathrm{MOR}}$: 0.588, AUC: 0,921.

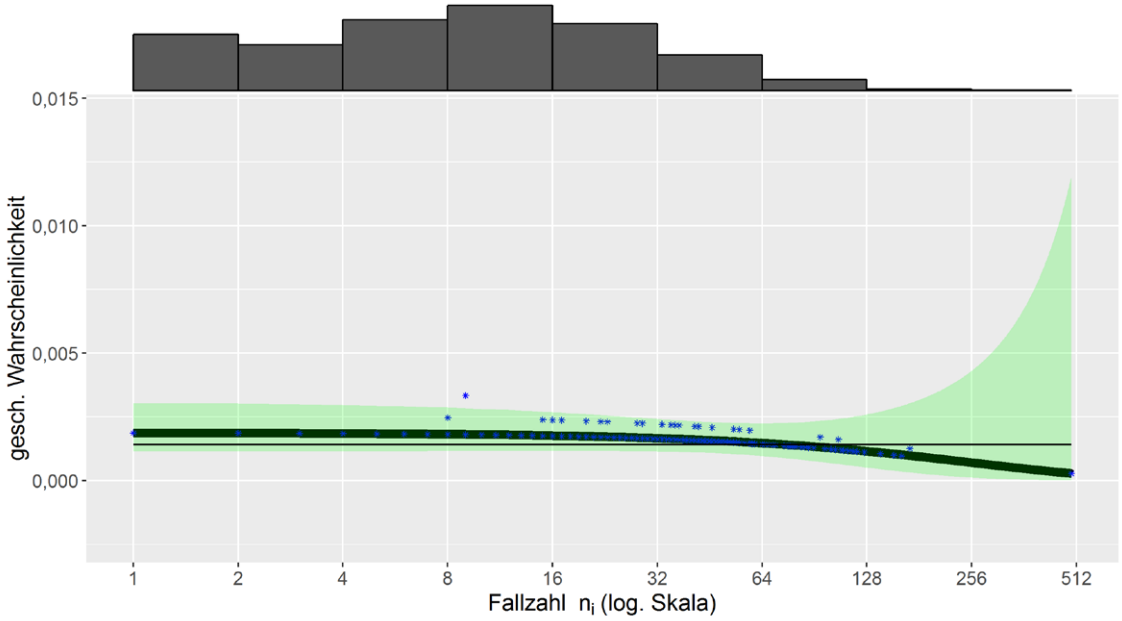

Abbildung 12: Verlaufsgrafik QI 2220 (Gefäßläsion)

Eine Schätzung dieser Form hat sich bereits bei anderen Qualitätsindikatoren ergeben.

Tabelle 16: Odds-Ratios QI 2220 (Gefäßläsion)

	Fallzahl n_1				
Fallzahl n_2	**4**	**12**	**24**	**52**	**104**
4	1	0,97	0,926	0,832	0,682
12		1	0,955	0,858	0,703
24			1	0,898	0,736
52				1	0,82
104					1

5.2.5 QI 51874: Postoperative Wundinfektionen ohne präoperative Infektzeichen

Tabelle 17: Zusammenfassung QI 51874 (Wundinfektion)

Zähler	Operationen, bei denen beim Patienten eine postoperative Wundinfektion auftrat.
Nenner	Alle Operationen bei Patienten ab 20 Jahren ohne Entzündungszeichen im Labor mit negativem Erregernachweis und aseptischem Eingriff.
Risikoadjustierung	Nicht vorhanden.

Statistische Ergebnisse	p-Wert: 0,2692, $\widehat{\mathrm{MOR}}$: 0,627, AUC: 0,837.

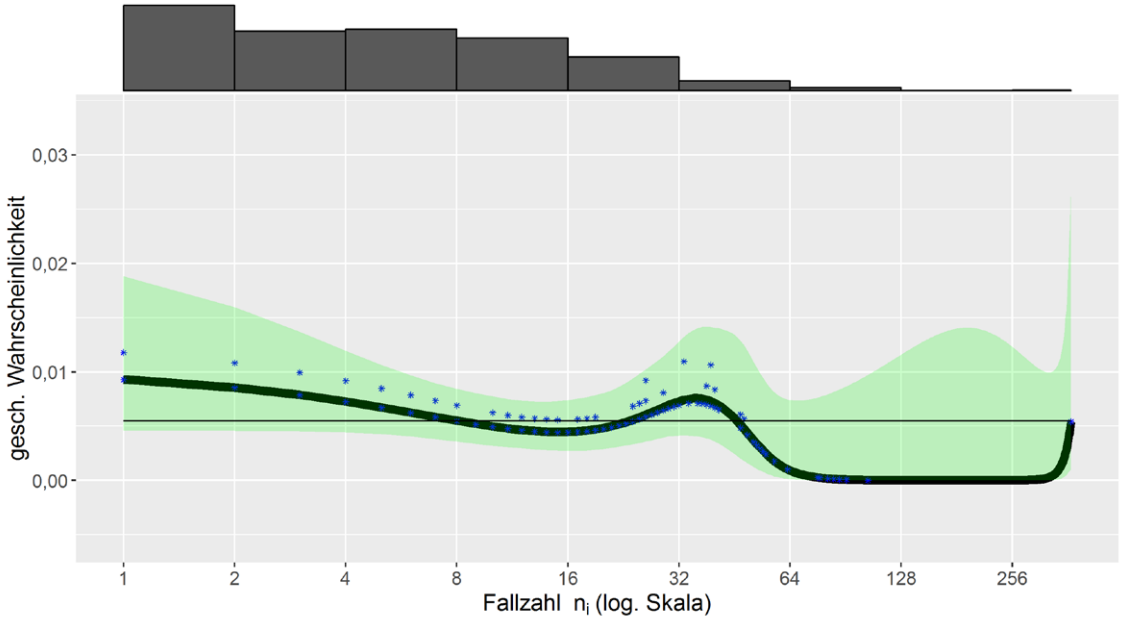

Abbildung 13: Verlaufsgrafik QI 51874 (Wundinfektion)

Der Verlauf der Kurve ist im Vergleich zu den anderen Ergebnissen sehr speziell. Zur Einordnung und als Hinweis, dies nicht überzubewerten, seien einige Aspekte genannt: Eine im Vergleich zu allen anderen Qualitätsindikatoren kleinere Datenbasis (Nennerbedingung), damit verbunden hohe Unsicherheit (Konfidenzbänder) und ein hoher *p*-Wert. Darüber hinaus war der qualitative Verlauf dieser Schätzung in unseren Sensitivitätsanalysen (siehe Abschnitt 5.4) relativ empfindlich – einen hohen *p*-Wert und die entsprechende statistische Bewertung erhalten wir jedoch für alle getesteten Konstellationen.

Bei diesem Qualitätsindikator wird die Variabilität der Grundkompetenzen als am geringsten geschätzt.

Tabelle 18: Odds-Ratios QI 51874 (Wundinfektion)

	Fallzahl n_1				
Fallzahl n_2	**4**	**12**	**24**	**52**	**104**
4	1	0,639	0,767	0,462	0,004
12		1	1,2	0,723	0,006
24			1	0,602	0,005
52				1	0,008
104					1

5.3 Überblick

Die Ergebnisse noch einmal in einer gemeinsamen Grafik:

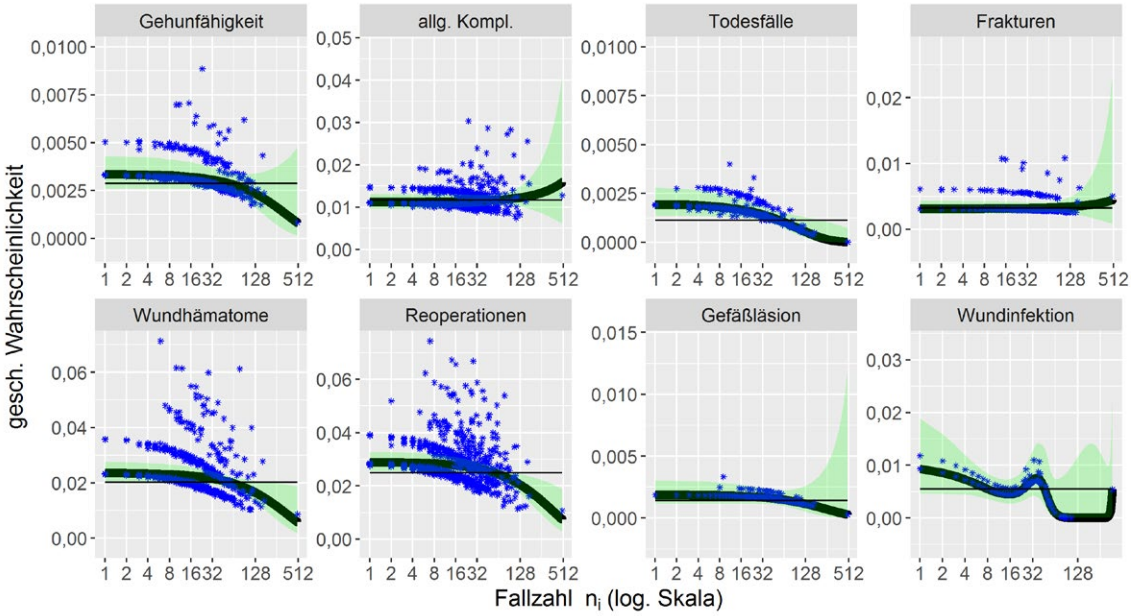

Abbildung 14: Verlaufsgrafiken im Überblick

An dieser Stelle sei noch einmal betont, dass alle Auswertungen für jeweils einen einzelnen Qualitätsindikator durchgeführt wurden.

5.4 Sensitivitätsanalysen

Spezifikation der Splines

Wie in den Abschnitten 4.2 und 4.3 erwähnt, gibt es unterschiedliche Möglichkeiten zur Wahl der Splines und Knoten.

Ein zentraler Aspekt ist dabei die Tatsache, dass die Fallzahlen sehr ungleichmäßig verteilt sind, sodass für die Schätzung von f in verschiedenen Fallzahlbereichen unterschiedlich viele Daten zur Verfügung stehen. Die klassische Herangehensweise, äquidistante Knoten mit gleichbleibender Glättung zu wählen, eignet sich in einer solchen Situation nicht: Einerseits fälllt die geschätzte Kurve in Bereichen mit vielen Daten tendenziell zu einfach aus und andererseits greifen die Glättungs- und Bestrafungsbedingungen in Bereichen (fast) ohne Daten so schlecht, dass die geschätzte Kurve unter sehr großer Unsicherheit unangemessen stark ausschlagen kann.

Um dieser Situation gerecht zu werden, existieren (mindestens) 3 Möglichkeiten:

1. Transformation der Fallzahl. Ersetzt man n_i durch zum Beispiel $\sqrt{n_i}$ oder $\mathrm{Log}(n_i)$, werden die Abstände zwischen den Fallzahlen und somit auch die Lücken in den Daten deutlich verkleinert, sodass keine großen Bereiche ohne Daten mehr vorkommen. Grundsätzlich kann dies zu sinnvollen Ergebnissen führen; es ist allerdings Vorsicht geboten, da dies in gewissem Sinne die tatsächliche Datenstruktur verschleiert und zum Beispiel zu Unterschätzung der statistischen Unsicherheit führen kann. Die Wahl der Transformation ist zudem sehr willkürlich. Die

beiden genannten typischen Varianten würden zum Beispiel auch die Fallzahlbereiche mit relativ vielen Daten (unnötig) verzerren.

2. Adaptive Glättung mit äquidistanten Knoten. Hierbei wird der Glättungsparameter λ lokal derart an die Anzahl der verfügbaren Daten angepasst, dass den beiden oben genannten Tendenzen entgegengewirkt wird. Der Ansatz ist wirksam, macht jedoch ein komplizierteres Modell mit zusätzlichen Parametern notwendig.
3. Quantilbasierte Knoten. Eine anderer Ansatz, um die beiden oben genannten Tendenzen zu kompensieren, besteht darin, die Knoten so zu platzieren, dass die einzelnen Teilstücke, auf denen f zunächst durch Polynome approximiert wird, gleich viele Daten enthalten. Auch dies ist wirksam und hat den Vorteil, dass weder die Daten verändert werden, noch das Modell komplizierter wird.

Das IQTIG hat alle drei Varianten geprüft (die erste Variante mit einer logarithmischen Transformation), wobei jeweils noch verschiedene Knotenzahlen zu wählen sind, und hat anhand der Maße AUC und AIC (siehe Abschnitt 4.3) sowie der genannten Vor-und Nachteile die Ansätze 2 und 3 für den vorliegenden Anwendungsfall als sinnvoll befunden. Die letztlich präsentierte Methodik (Abschnitt 4.2) und die obigen Ergebnisse beruhen auf Ansatz 3, da dieser grundsätzlich einfacher ist als Ansatz 2 und äquivalente Ergebnisse liefert.

Es sei an dieser Stelle betont, dass die Wahl des Ansatzes keinen Einfluss auf die Signifikanzeinstufung hat: Der statistische Test lehnt stets bei den gleichen drei Qualitätsindikatoren zum Niveau α = 5 % die Nullhypothese ab.

Einfluss von Ausreißern

Da es bei allen Qualitätsindikatoren jeweils einen Standort mit besonders hoher Fallzahl gibt (396 bzw. 495 Fälle), der einen entsprechend großen Beitrag zur ungleichmäßigen Fallzahlverteilung liefert, haben wir als weitere Sensitivitätsanalyse die Berechnungen unter Ausschluss dieses Ausreißers durchgeführt. Es zeigt sich, dass dies weder einen Einfluss auf die statistischen Signifikanzeinstufungen hat, noch auf den qualitativen Kurvenverlauf bei den Qualitätsindikatoren mit statistisch signifikantem Fallzahleffekt.

Vergleich mit auf Fallzahlgruppen basierenden Auswertungsmethoden

Es wurden Sensitivitäts-Analysen mit auf Fallzahlgruppen basierenden Methoden durchgeführt. Wie in Abschnitt 4.2 diskutiert, hängen diese Ergebnisse zum Teil erheblich von der spezifischen Wahl der Fallzahlgruppen ab. Die qualitative Aussage eines unterhalb des Bundesdurchschnitts liegenden Ergebnisses in der jeweils größten Fallzahlgruppe lässt sich jedoch auch mit dieser Methodik für die drei Qualitätsindikatoren bestätigen, die auch auf Basis des Inferenzmodells (1) signifikante Zusammenhänge zwischen Fallzahl und Ergebnis zeigten.

6 Diskussion

Die vorliegende retrospektive Analyse des Leistungsbereichs „Knie-Endoprothesenwechsel und -komponentenwechsel“ basiert auf 17.685 Wechselprozeduren, die in 1.085 Standorten dokumentiert wurden. Deskriptiv zeigt die Verteilung der Fallzahl auf die Standorte eine hohe Zahl an Krankenhäusern, die lediglich 1 bis 2 Fälle pro Jahr operieren. Übertragen auf den Anteil aller Wechselprozeduren wird ca. die Hälfte aller Wechselprozeduren in Standorten mit ca. 16 bis 64 Fällen durchgeführt (vgl. Abbildung 2. Seite 18).

Die zu untersuchenden Qualitätsindikatoren weisen überwiegend eine niedrige Prävalenz auf. Vergleichsweise häufig treten die interessierenden Ereignisse der Indikatoren „Reoperationen aufgrund von Komplikationen“ (3,7 %), „Allgemeine postoperative Komplikationen“ (1,9 %) und „Wundhämatome/Nachblutungen“ (2,2 %) auf. Aus ihrer geringen Prävalenz ergibt sich ein hoher Anteil an Standorten, die keine Fälle mit interessierendem Ereignis in den einzelnen Ergebnisindikatoren haben (vgl. Tabelle 2, Seite 19).

In Zusammenschau wird für drei der acht Indikatoren ein statistisch signifikanter Fallzahl-Ergebnis-Zusammenhang gefunden. Für die drei Indikatoren „Wundhämatome/Nachblutung“, „Todesfälle während des akut stationären Aufenthaltes“ und „Reoperation aufgrund von Komplikationen“ wird mit steigender Fallzahl des Krankenhauses eine sinkende Wahrscheinlichkeit für das Outcome geschätzt. Für die Indikatoren „Gehunfähigkeit“, „Allgemeine Komplikationen“, „Frakturen“, „Gefäßläsion“ und „Infektion“ wird kein statistisch signifikanter Zusammenhang zwischen Auftreten des jeweiligen interessierenden Ereignisses und der Leistungsmenge des Krankenhauses gefunden. Innerhalb der Spannweite der Konfidenzbänder, die das Spektrum mit den Daten „in Einklang zu bringender“ Fallzahleffekte abbilden, können für diese übrigen fünf Indikatoren sowohl steigende als auch fallende und sogar nicht monotone Verläufe des Fallzahleffektes beobachtet werden.

Die Ergebnisse sollen anhand eines Beispiels verdeutlicht werden: Unter Annahme einer durchschnittlichen fallzahlunabhängigen Kompetenz des Standortes und einem durchschnittlichen Risiko eines Falles kann ein signifikanter Zusammenhang zwischen der Leistungsmenge und der Wahrscheinlichkeit für eine „Reoperation aufgrund von Komplikationen“ festgestellt werden. Entsprechend wird für einen Standort mit sehr wenigen Fällen eine Wahrscheinlichkeit von 2 bis 3 % geschätzt, wohingegen diese für den größten Standort auf ca. 1 % geschätzt wird. Diese Schätzung ist mit einer sehr hohen statistischen Unsicherheit behaftet, da die Schätzung insbesondere im hohen Fallzahlbereich auf wenigen Standorten basiert, und im Bereich niedriger Fallzahlen nur wenige Behandlungsfälle in die Schätzung einfließen (vgl. Abbildung 1, Seite 17). Die grafische Darstellung des Zusammenhangs zeigt zudem, dass dieser Zusammenhang nicht linear anzunehmen ist (vgl. Abbildung 7, Seite 39). Während die geschätzte Wahrscheinlichkeit für eine Reoperation im Bereich kleiner bis mittlerer Fallzahlen nahezu gleich ist, zeigt sich ab einer mittleren bis hohen Fallzahl ein Abfall.

Die Variabilität auf Ebene der Standorte ist sehr groß. So unterscheidet sich die geschätzte Wahrscheinlichkeit für eine „Reoperation aufgrund von Komplikationen“ für einen Durchschnittspatienten bzw. eine Durchschnittspatientin bei Behandlung in den einzelnen Standorten deutlich und liegt in seinen Extremen zwischen 7 bis 8 % und 1 %. Deutlich wird zudem, dass für einzelne Standorte mit sehr wenigen behandelten Fällen eine deutlich geringere Wahrscheinlichkeit für das Auftreten des interessierenden Ereignisses geschätzt wird. Umgekehrt gibt es sehr große Standorte für die sich eine sehr große Ereigniswahrscheinlichkeit im untersuchten Qualitätsindikator ergibt. Für eine Patientin oder einen Patienten mit durchschnittlichem Risiko liegt die Wahrscheinlichkeit für das Auftreten einer Reoperation im Durchschnitt aller Standorte bei 2,5 %.

Ähnlich lassen sich die Ergebnisse für den Qualitätsindikator „Wundhämatome/Nachblutungen“ und „Todesfälle“ darstellen. Allerdings basieren letztere auf weniger Fällen im interessierenden Ereignis (Prävalenz von 0,4 %). Für das Auftreten allgemeiner Komplikationen wird bei vergleichsweise hoher Prävalenz (1,9 %) kein statistisch signifikanter Volumen-Outcome-Zusammenhang gefunden. Hervorgehoben werden soll an dieser Stelle, dass aus einer statistischen Signifikanz keine klinische Relevanz abgeleitet werden kann.

Limitationen

Um oben genannte Ergebnisse hinsichtlich eines klinisch relevanten Zusammenhangs zwischen Volumen und Outcome bei Knieendoprothesen-Wechseleingriffen zu bewerten, müssen die Limitationen der vorliegenden Analyse näher betrachtet werden. Hier soll zunächst auf die verfügbare Datenbasis hingewiesen werden. Diese beruht auf den Ergebnissen *eines* Erfassungsjahres, was in Kombination mit der geringen Prävalenz der Outcome-Parameter dazu führt, dass die absolute Anzahl interessierender Ereignisse in den Indikatoren äußert gering ist (vgl. Tabelle 2). Infolge dieser eingeschränkten Datenbasis ist nur eine gewisse Effektstärke messbar. Mit dem gewählten Signifikanzniveau von 5 % konnte lediglich für drei der acht Indikatoren ein signifikanter Fallzahl-Ergebnis-Zusammenhang gefunden werden. Es ist durchaus möglich, dass sich statistisch signifikante Zusammenhänge auch für die anderen Indikatoren finden ließen, wenn eine größere Datenbasis zugrunde läge. Signifikante Fallzahleffekte müssten jedoch auch dann noch auf ihre Relevanz für die Versorgungssituation hin überprüft werden.

Aus Versorgungssicht ist die geringe Prävalenz der Outcomes der Ergebnisindikatoren sicherlich gut zu bewerten, aus biometrischer Sicht ergibt sich jedoch eine hohe statistische Unsicherheit der modellbasierten Schätzungen und somit eine eingeschränkte Übertragbarkeit auf die tatsächliche Beurteilung der Versorgungssituation. Eine Erweiterung der Datenbasis wäre durch eine Erfassungsjahr-übergreifende Auswertung möglich gewesen und hätte ggf. eine höhere statistische Aussagekraft der Ergebnisse ermöglicht. Allerdings wurde im Erfassungsjahr 2014 auf eine standortbasierte Erhebung und Auswertung umgestellt, sodass sich die zugrunde liegenden Auswertungseinheiten zwischen den Erfassungsjahren unterscheiden (vgl. Kapitel 3, Seite 20). Ein valider Vergleich der Fallzahlen als möglicher Surrogatparameter für die Kompetenz eines Krankenhauses wäre aufgrund der verschiedenartigen Struktur- und Prozessmerkmale unzulässig, sodass in dieser Sonderauswertung von der Erweiterung der Datenbasis abgesehen wurde.

Darüber hinaus ist die Auswertung auf Standortebene an sich als Limitation zu bewerten (Luft et al. 1987). Ein Standort kann verschiedene Fachabteilungen einschließen. Knieendoprothesen-Wechsel werden sowohl von allgemein-chirurgischen, unfallchirurgischen als auch von orthopädischen Fachabteilungen durchgeführt. Diese können an großen Standorten parallel existieren, werden in dieser Sonderauswertung jedoch aus Mangel an präziseren Informationen als eine Einheit betrachtet. Im vorliegenden Analysemodell wird daher für alle Fachabteilungen innerhalb eines Standortes eine einheitliche Qualität unterstellt. Mögliche Clusterbildungen innerhalb eines Standortes können nicht abgebildet werden. Zudem werden vorliegende Analysen auf Basis der entlassenden und nicht der behandelnden Standorte durchgeführt. Im Einzelfall kann die Ursache für den Eintritt eines Ereignisses somit nicht dem dokumentierenden Standort zugeschrieben werden, sondern liegt im Verantwortungsbereich des behandelnden Standortes.

Die zeitliche Begrenzung der fallbasierten QS-Dokumentation auf den stationären Aufenthalt limitiert die Übertragbarkeit der dargestellten Ergebnisse erheblich. Die beauftragten Indikatoren schließen lediglich alle im Rahmen des Krankenhausaufenthaltes erfassten Outcome-Parameter ein. Sterbefälle oder Komplikationen im poststationären Zeitraum können auf Basis von QS-Daten derzeit nicht erfasst werden. Dies ist für die Beurteilung der vorliegenden Analyse von entscheidender Bedeutung. Taylor et al. (1997) berichten für Knieendoprothesenwechsel eine intra-hospitale Mortalität von 0,28 %, wohingegen eine 30-Tages-Mortalität von 0,48 % angegeben wird. Eine sehr hohe Zahl an nicht erfassten Todesfällen, die jedoch mit dem operativen Eingriff assoziiert sein können, ist daher in vorliegender Analyse anzunehmen. Die Hinzunahme von Sozialdaten der Krankenkassen bzw. Follow-up-Indikatoren wären Möglichkeiten, die Datenbasis dahingehend zu erweitern.

Ebenfalls limitierend ist, dass die vorliegenden QS-Daten aus einer Beobachtungsstudie stammen. Dies erlaubt lediglich den Nachweis von Assoziationen, lässt jedoch keinen Rückschluss auf mögliche kausale Zusammenhänge zu. Einschränkend ist hinzuzufügen, dass die Erhebung der Daten nicht zum Zweck der Darstellung von Volumen-Outcome-Analysen erfolgte. Vielmehr wurden die Daten zum Zweck der externen stationären Qualitätssicherung erhoben und basieren damit auf einer Selbstdokumentation der beteiligten Standorte. Dem Interesse vollumfänglicher Informationen steht somit immer das Argument der Datensparsamkeit gegenüber, was die geringe Verfügbarkeit patientenseitiger Risikofaktoren begründet. Zwar wurden die Risikoadjustierungen der Qualitätsindikatoren für die vorliegende Analyse übernommen, diese bilden das patientenseitige Risikoprofil für Knieendoprothesenwechsel jedoch unvollständig ab. Relevante Störvariablen, die das Outcome in der Knieendoprothetik beeinflussen und in der Qualitätssicherung erhoben werden, sind u. a. Alter, Geschlecht und ASA-Klassifikation der Patientin bzw. des Patienten. Darüber hinaus existieren weitere Confounder, die im Rahmen der Qualitätssicherung jedoch nicht erhoben werden. Dazu gehören z. B. kardiovaskuläre Nebenerkrankungen und die Dringlichkeit der Krankenhausaufnahme. Diese sind in der vorliegenden Analyse nicht bekannt, können aber entscheidend für das Outcome in den einzelnen Indikatoren sein (Wetzel 2006).

Den patientenseitigen Risikofaktoren gesellt sich eine Vielzahl nicht erfasster Merkmale auf Seite der Leistungserbringer hinzu. Vor dem Hintergrund der „practice-makes-perfect"-Hypothese wäre ein Wissen um die Erfahrung der Operateurin bzw. des Operateurs bzw. des gesamten OP-Teams eine wichtige Information (Schräder et al. 2007). Entsprechende Analysen liegen für die Primärendoprothetik vor und die Erkenntnisse können für die Wechseleingriffe ebenfalls angenommen werden (Varagunam et al. 2015). Wichtige personelle und strukturelle Merkmale, die die Kompetenz des Leistungserbringers möglicherweise besser abbilden, sind nicht bekannt. Informationen über die Verfügbarkeit und die Nutzung relevanter Strukturvariablen (z. B. Computer Tomographie) oder Prozessvariablen (z. B. klinische Behandlungspfade) liegen nicht vor. Auch der Einfluss des Behandlungsteams lässt sich nicht einschätzen. Dies schließt das perioperative Versorgungsmanagement ebenso mit ein, wie pflegerische und physiotherapeutische Maßnahmen in den Tagen nach der Operation. Ihr Einfluss auf die Volumen-Outcome-Beziehungen i. S. eines Confounders kann folglich mit der vorliegenden Analyse nicht eingeschätzt werden.

Ebenfalls zu beachten ist die Auswahl der für die Durchführung der Qualitätssicherung zu dokumentierenden Fälle. Die externe stationäre Qualitätssicherung erfasst nicht sämtliche Knieendoprothesenwechsel deutschlandweit. Einige zumeist fachlich sehr komplexe Fälle sind von der Dokumentationspflicht ausgeschlossen. Beispielhaft seien hier bösartige Neubildungen an Knorpel- oder Knochengewebe genannt. Rückschlüsse auf Volumen-Outcome-Zusammenhänge für diese besonderen Fallkonstellationen können somit nicht getroffen werden.

Betrachtet man die Vollständigkeit aller dokumentationspflichtigen Fälle, kommt hinzu, dass nicht sicher gesagt werden kann, ob alle Outcome-Parameter (z. B. Komplikationen) vollständig und korrekt dokumentiert wurden. Dies soll am Beispiel der Datenvalidierung des Erfassungsjahres 2017 verdeutlicht werden (IQTIG 2019a). Hier wurde für das Datenfeld „OP oder interventionsbedürftige/-s Nachblutung/Wundhämatom" auf Basis einer Stichprobe eine Übereinstimmungsrate zwischen QS-Dokumentation und Patientenakte von lediglich 71 % festgestellt. Dies limitiert die Aussagekraft vorliegender Sonderauswertung, da mögliche Auswirkungen oder Zusammenhänge mit der Leistungsmenge nicht ausgeschlossen werden können.

Unter Annahme der „practice-makes-perfect"-Hypothese könnte zudem vermutet werden, dass „low-volume"-Standorte etwaige Komplikationen seltener sehen und somit seltener erkennen bzw. diagnostizieren (Luft et al. 1987). Die Differenz zwischen tatsächlich vorliegenden und diagnostizierten (und damit auch dokumentierten) Ereignissen würde sich zwischen Standorten mit hoher und geringer Leistungsmenge unterscheiden und die Volumen-Outcome-Beziehung zuungunsten der „high-volume"-Standorte verändern. Zudem könnten einem Krankenhaus mit Zentrumscharakter häufiger komplexere Fälle überwiesen werden, die in kleineren Krankenhäusern nicht behandelt werden können. Eine derartige Ungleichheit im modellierbaren Risikoprofil der Patientinnen und Patienten hätte eine Verzerrung der Ergebnisse zugunsten der kleinen Standorte zur Folge.

Demgegenüber stehen jedoch auch Faktoren, die die Risikoadjustierung zugunsten größerer Krankenhäuser verzerren könnten: I. S. eines „cream-skimming" ist zum Beispiel nicht auszuschließen, dass aufgrund der Vielzahl an möglichen Patientinnen und Patienten eine Vorauswahl

„leichterer“ Fälle stattfindet. Niedrigere Raten an Reoperationen oder seltenere Todesfälle wären durch diesen Effekt zumindest teilweise erklärbar.

Ebenfalls unklar bleibt in vorliegender Analyse die Spezialisierung der Standorte auf besondere Patientengruppen. Möglicherweise behandeln einzelne Krankenhäuser lediglich ausgewählte Patientengruppen (z. B. Rheumatiker) und verfügen für diese Patientengruppe über eine sehr hohe Kompetenz. Aufgrund der Seltenheit der Kombination aus Patientengruppe und Knieendoprothesenwechsel würde das entsprechende Krankenhaus im oben genannten Fall nicht zu denjenigen Standorten gezählt, die eine hohe Leistungsmenge erbringen.

Auch auf Ebene der gewählten Eingriffsart ist eine Spezialisierung auf Standortebene wahrscheinlich. So wird im Jahresbericht des Endoprothesenregisters 2019 festgestellt, dass viele Krankenhäuser lediglich zu einem geringen Anteil unikondyläre Schlittenprothesen einsetzen (Grimberg et al. 2019). Gleichzeitig existieren einige wenige Krankenhäuser, die fast ausschließlich Unischlitten im Rahmen einer Primärversorgung implantieren. Eine ähnliche Spezialisierung ist auch für den Knieendoprothesenwechsel anzunehmen. Da Informationen dazu nicht vorliegen, konnte dies in der vorliegenden Analyse nicht berücksichtigt werden. Diese Spezialisierung impliziert bereits, dass aus medizinischer Sicht verschiedene – unter dem Überbegriff „Knie-Endoprothesenwechsel bzw. -komponentenwechsel“ zusammengefasste – Eingriffe mit einer unterschiedlichen Komplikationsrate einhergehen (vgl. Kapitel 1). Aus fachlicher Sicht ist dies leicht nachvollziehbar: so ist der einfache Inlaywechsel im Vergleich zur Entfernung und Reimplantation knochenverankerter Komponenten deutlich schwieriger einzustufen. Gleiches ist für einen zunehmenden zeitlichen Abstand zwischen Ersteingriff und Revision der Knieendoprothese anzunehmen. Die Folge sind variierende Anforderungen an die Kompetenz des Leistungserbringers, die in der vorliegenden Analyse nicht differenziert betrachtet werden können. Ebenfalls nicht unterschieden werden konnte hinsichtlich der Indikation des Wechseleingriffs. Die häufigste Indikation ist die Lockerung des Primärimplantats. Diese kann auf einer Infektion oder mechanischem Verschleiß bzw. Defekt basieren. Für die Operationsplanung ist diese Unterscheidung außerordentlich wichtig, da das Risikoprofil des Eingriffs möglicherweise deutlich variiert. Obwohl entsprechende Subgruppen in den QS-Daten identifizierbar wären und grundsätzlich in Form sogenannter Interaktionen im Modell bedacht werden könnten, wurden derartige Subgruppenanalysen nicht durchgeführt, da die statistische Unsicherheit einer derartigen Analyse in Anbetracht der Seltenheit der analysierten Outcome-Parameter sehr groß wäre. Durch eine retrospektive Poweranalyse könnte simulationsbasiert quantifiziert werden, wie groß die Effektstärke etwaiger Fallzahl-Ergebnis-Zusammenhänge sein müsste, um bei der vorliegenden Datengrundlage überhaupt detektiert werden zu können. Die breiten Konfidenzbänder zeigen jedoch bereits im Modell ohne Interaktionen, dass die Unsicherheit zu groß ist, um aus den Subgruppenanalysen substanzielle Ergebnisse zu gewinnen. Da die Datengrundlage im Rahmen dieses Auftrages ohnehin nicht sinnvoll erweiterbar war (vgl. Kapitel 3), wurde von einer solchen retrospektiven Poweranalyse abgesehen.

Die Übertragbarkeit der vorliegenden Analyse auf die klinische Versorgungsrealität wird durch eine weitere Annahme des Analyse-Modells limitiert. Die kumulative Auswertung der Daten eines Erfassungsjahres setzt voraus, dass die Kompetenz des Leistungserbringers zeitlich konstant

ist. Hypothetisch könnte beispielsweise das Auftreten einer schweren Komplikation oder anderer äußerer Umstände zu einer Anpassung personeller, struktureller oder prozeduraler Gegebenheiten im Standort führen und unterjährig die Kompetenz des Leistungserbringers beeinflussen. Diese können im Einzelfall die Änderung der Ergebnisqualität zum Ziel haben und wirken sich daher auf die Fallzahl-Ergebnis-Beziehung aus. In den QS-Daten sind eine Chronologie der Behandlungsfälle bzw. unterjährige Anpassungen der Versorgungsqualität nicht darstellbar (vgl. Abschnitt 4.1, dort Seite 22).

7 Fazit und Empfehlungen

Die vorliegende Sonderauswertung liefert Hinweise auf eine Beziehung zwischen Leistungsmenge und Behandlungsergebnis – ein versorgungsrelevanter Zusammenhang kann durch die vorliegende Analyse jedoch nicht belegt werden. Zwar werden in einzelnen Qualitätsindikatoren statistisch signifikante Zusammenhänge zwischen Volumen und Outcome festgestellt, allerdings trifft dies lediglich für drei der acht untersuchten Qualitätsindikatoren zu. Zur Beurteilung der Aussagekraft für die Versorgungssituation sollte auch die geringe Prävalenz der Outcomes beachtet werden. Die Zusammenhangsanalysen wurden unabhängig für die verschiedenen Indikatoren durchgeführt. Insbesondere da für die unterschiedlichen Indikatoren verschiedene (teilweise nicht signifikante) Zusammenhänge festgestellt wurden, bedürfte es zukünftig einer Gewichtung der Qualitätsindikatoren entsprechend ihrer Relevanz für die Versorgungssituation.

Es ist in vorliegender Analyse nicht trennbar, ob die beobachteten Zusammenhänge sich durch eine fallzahlabhängige Ausprägung nicht beobachteter patientenseitiger Risikofaktoren (ggf. sogar durch eine gezielte fallzahlabhängige Patientenselektion seitens der Leistungserbringer (sog. „cream-skimming“)) oder durch einen tatsächlich bestehenden Zusammenhang von Leistungserbringer-Kompetenz und Fallzahl erklären. Unter Berücksichtigung der vielschichtigen Limitationen entsteht so eine Wissenslücke zwischen den Ergebnissen der vorliegenden Analyse und dem Ziel einer Darstellung der realen Versorgungsqualität in Abhängigkeit der erbrachten Leistungsmenge.

Um diese weiter zu schließen bedarf es Anstrengungen auf verschiedenen Ebenen: Eine Stärkung der vorliegenden Analyse wäre beispielsweise durch den Einbezug einer größeren Datengrundlage möglich. Die Auswertung der Ergebnisse mehrerer Jahre bei gleichbleibender Auswertungseinheit und Berechnung der Ergebnisindikatoren würde die Analyse auf eine breitere Basis stellen. Denkbar wäre auch, weitere Datenquellen wie z. B. Sozialdaten bei den Krankenkassen einzuschließen, um möglicherweise erst poststationär auftretende Komplikationen erfassen zu können. Eine präzisere Eingrenzung der Eingriffsart würde die inhaltliche Aussagekraft darüber hinaus stärken. Sofern entsprechende Informationen vorhanden sind, wäre die Berücksichtigung von Subgruppen (z. B. nach Indikationen) möglich oder eine Aussage darüber, ob es sich im Einzelfall um Revisionseingriffe nach Ersteingriffen im selben Krankenhaus handelt.

In der statistischen Modellierung sind Weiterentwicklungen denkbar, die gezielt den Einfluss weiterer Unsicherheitsquellen untersuchen: So wird beispielsweise in der vorliegenden Analyse keine Schätzungssicherheit der von AQUA (2015b) übernommenen Risikoadjustierung berücksichtigt. Des Weiteren ist eine andere Modellierung der Random-Effects und deren A-priori-Verteilung denkbar, etwa durch Abweichung von der in diesen Analysen gewählten Normalverteilungsannahme. Um mögliche Einflussfaktoren auf Seite der Patientinnen und Patienten und auf Seite der Leistungserbringer genauer berücksichtigen zu können, müssten Outcome-spezifische Risikofaktoren sowie personelle, strukturelle und prozedurale Voraussetzungen der Krankenhäuser erhoben werden. Prospektiv randomisierte Studien sind für die gewählte Fragestellung

nur bedingt umsetzbar, sodass quasi-experimentelle Studiendesigns in Betracht gezogen werden sollten. Diese könnten Antworten auf kausale Zusammenhänge zwischen Kompetenz des Leistungserbringers i. S. des Behandlungsergebnisses und der erbrachten Leistungsmenge liefern. Generell zeigt sich somit, dass die Erforschung von Volumen-Outcome-Zusammenhängen eine grundlegende Auseinandersetzung mit der Frage erfordert, welche methodischen Standards an den Nachweis solcher Zusammenhänge gesetzt werden sollen, um evidenzbasierte gesundheitspolitische Schlussfolgerungen ziehen zu können.

Literatur

Agresti, A (2013): Categorical Data Analysis. 3. Auflage. Hoboken, New Jersey: Wiley-Interscience. ISBN: 978-0-47-046363-5.

AQUA [Institut für angewandte Qualitätsförderung und Forschung im Gesundheitswesen] (2015a): 17/7 – Knie-Endoprothesenwechsel und -komponentenwechsel. Bundesauswertung zum Erfassungsjahr 2014. Erstellt am: 19.05.2015. Göttingen: AQUA. 24/2015020001. URL: https://sqg.de/downloads/Bundesauswertungen/2014/bu_Gesamt_17N7-KNIE-WECH_2014.pdf (abgerufen am: 11.12.2019).

AQUA [Institut für angewandte Qualitätsförderung und Forschung im Gesundheitswesen] (2015b): Knie-Endoprothesenwechsel und -komponentenwechsel. Beschreibung der Qualitätsinidkatoren für das Erfassungsjahr 2014. Stand: 05.05.2015. Göttingen: AQUA. URL: https://sqg.de/downloads/QIDB/2014/AQUA_17n7_Indikatoren_2014.pdf (abgerufen am: 14.10.2019).

Bachmann, MO; Alderson, D; Edwards, D; Wotton, S; Bedford, C; Peters, TJ; et al. (2002): Cohort study in South and West England of the influence of specialization on the management and outcome of patients with oesophageal and gastric cancers. *British Journal of Surgery* 89(7): 914-922. DOI: 10.1046/j.1365-2168.2002.02135.x.

Bender, R; Grouven, U (2006): Möglichkeiten und Grenzen statistischer Regressionsmodelle zur Berechnung von Schwellenwerten für Mindestmengen. *Zeitschrift für Ärztliche Fortbildung und Qualität im Gesundheitswesen* 100(2): 93-98.

Birkmeyer, JD; Siewers, AE; Finlayson, EVA; Stukel, TA; Lucas, FL; Batista, I; et al. (2002): Hospital Volume and Surgical Mortality in the United States. *The New England Journal of Medicine* 346(15): 1128-1137. DOI: 10.1056/NEJMsa012337.

Bishop, CM (2006): Pattern Recognition And Machine Learning. (Information Science and Statistics). New York, US-NY: Springer Science+Business Media. ISBN: 978-0387-31073-2. URL: https://www.microsoft.com/en-us/research/uploads/prod/2006/01/Bishop-Pattern-Recognition-and-Machine-Learning-2006.pdf (abgerufen am: 09.10.2019).

Fahrmeir, L; Kneib, T; Lang, S (2009): Regression. Modelle, Methoden und Anwendungen. Zweite Auflage. (Statistik und ihre Anwendungen). Heidelberg [u. a.]: Springer. ISBN: 978-3-642-01836-7.

Feinglass, J; Koo, S; Koh, J (2004): Revision Total Knee Arthroplasty Complication Rates in Northern Illinois. *Clinical Orthopaedics and Related Research* 429: 279-285. DOI: 10.1097/01.blo.0000137563.27841.e9.

Fitzmaurice, GM; Laird, NM; Ware, JH (2011): Applied Longitudinal Analysis. Second Edition. (Wiley Series in Probability and Statistics). Hoboken, US-NJ [u. a.]: John & Wiley Sons. ISBN: 978-0-470-38027-7.

G-BA [Gemeinsamer Bundesausschuss] (2019): Beschluss des Gemeinsamen Bundesausschusses über eine Beauftragung des IQTIG mit der Auswertung von esQS-Daten hinsichtlich Volume-Outcome-Beziehungen bei Revisionseingriffen bei Knie-Endoprothesen. [Stand:] 16.05.2019. Berlin: G-BA. URL: https://www.g-ba.de/downloads/39-261-3790/2019-05-16_IQTIG-Beauftragung_Auswertung-esQS-Daten-Knie-Endoprothesen.pdf (abgerufen am: 28.10.2019).

Gelman, A; Hill, J (2007): Data analysis using regression and multilevel/hierarchical models. 1. Auflage. Cambridge [u. a.]: Cambridge University Press. ISBN: 978-0-52-168689-1.

George, EI; Ročková, V; Rosenbaum, PR; Satopää, VA; Silber, JH (2017): Mortality Rate Estimation and Standardization for Public Reporting: Medicares' Hospital Compare. *Journal of the American Statistical Association* 112(519): 933-947. DOI: 10.1080/01621459.2016.1276021.

Ghaferi, AA; Birkmeyer, JD; Dimick, JB (2009): Complications, Failure to Rescue, and Mortality With Major Inpatient Surgery in Medicare Patients. *Annals of Surgery* 250(6): 1029-1034. DOI: 10.1097/SLA.0b013e3181bef697.

Grimberg, A; Jansson, V; Melsheimer, O; Steinbrück, A (2019): Jahresbericht 2019. Mit Sicherheit mehr Qualität. Berlin: EPRD [Deutsche Endprothesenregister]. ISBN: 978-3-9817673-4-6. URL: https://www.eprd.de/fileadmin/user_upload/Jahresbericht_2019_doppelseite_2.0.pdf (abgerufen am: 20.11.2019).

Grouven, U; Küchenhoff, H; Schräder, P; Bender, R (2008): Flexible regression models are useful tools to calculate and assess threshold values in the context of minimum provider volumes. *Journal of Clinical Epidemiology* 61(11): 1125-1131. DOI: 10.1016/j.jclinepi.2007.11.020.

Hosmer, DW; Lemeshow, S; Sturdivant, RX (2013): Applied Logistic Regression. Hoboken, US-NJ: Wiley. ISBN: 978-0-470-58247-3.

IQTIG [Institut für Qualitätssicherung und Transparenz im Gesundheitswesen] (2019a): Bericht zur Datenvalidierung 2018 (nach QSKH-RL). Erfassungsjahr 2017. Stand: 28.05.2019. Berlin: IQTIG. [unveröffentlicht].

IQTIG [Institut für Qualitätssicherung und Transparenz im Gesundheitswesen] (2019b): Methodische Grundlagen V1.1. Stand: 15.04.2019. Berlin: IQTIG. URL: https://iqtig.org/dateien/dasiqtig/grundlagen/IQTIG_Methodische-Grundlagen-V1.1_barrierefrei_2019-04-15.pdf (abgerufen am: 05.12.2019).

IQWIG [Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen] (2006): Entwicklung und Anwendung von Modellen zur Berechnung von Schwellenwerten bei Mindestmengen für die Koronarchirurgie. Abschlussbericht. Stand: 20.06.2006. (IQWIG-Berichte • Jahr: 2006 Nr. 9). Berlin: IQWIG. Auftrag B05/01b. URL: https://iqwig.de/de/projekte-ergebnisse/projekte/medizinische-biometrie/b05-01b-berechnung-von-schwellenwerten-

bei-mindestmengen-fuer-die-koronarchirurgie.1217.html [Aktuelles Dokument: Abschlussbericht > Download] (abgerufen am: 25.09.2019).

IQWIG [Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen] (2019): Zusammenhang zwischen Leistungsmenge und Qualität des Behandlungsergebnisses bei Lebertransplantation (inklusive Teilleber-Lebendspende). Stand: 04.09.2019. (IQWIG-Berichte – Nr. 813). Köln: IQWIG. Rapid Report V18-04. URL: https://www.iqwig.de/de/projekte-ergebnisse/projekte/versorgung/v18-04-zusammenhang-zwischen-leistungsmenge-und-qualitaet-des-behandlungsergebnisses-bei-lebertransplantationen-rapid-report.10904.html [Berichtsdokumente > Rapid Report > Dokument herunterladen] (abgerufen am: 18.10.2019).

Kizer, KW (2003): The Volume-Outcome Conundrum. *The New England Journal of Medicine* 349(22): 2159-2161. DOI: 10.1056/NEJMe038166.

Larsen, K; Petersen, JH; Budtz-Jørgensen, E; Endahl, L (2000): Interpreting Parameters in the Logistic Regression Model with Random Effects. *Biometrics* 56(3): 909-914. DOI: 10.1111/j.0006-341X.2000.00909.x.

Lee, Y; Nelder, JA (2004): Conditional and Marginal Models: Another View. *Statistical Science* 19(2): 219-238. DOI: 10.1214/088342304000000305.

Luft, HS; Bunker, JP; Enthoven, AC (1979): Should Operations be Regionalized? *The New England Journal of Medicine* 301(25): 1364-1369. DOI: 10.1056/nejm197912203012503.

Luft, HS; Hunt, SS; Maerki, SC (1987): The Volume-Outcome Relationship: Practice-Makes-Perfect or Selective-Referral Patterns? *Health Services Research* 22(2): 157-182. URL: https://europepmc.org/backend/ptpmcrender.fcgi?accid=PMC1065430&blobtype=pdf (abgerufen am: 18.11.2019).

Marra, G; Wood, SN (2012): Coverage Properties of Confidence Intervals for Generalized Additive Model Components. *Scandinavian Journal of Statistics* 39(1): 53-74. DOI: 10.1111/j.1467-9469.2011.00760.x.

Muff, S; Held, L; Keller, LF (2016): Marginal or conditional regression models for correlated non-normal data? *Methods in Ecology and Evolution* 7(12): 1514-1524. DOI: 10.1111/2041-210x.12623.

Pieper, D; Mathes, T; Neugebauer, E; Eikermann, M (2013): State of Evidence on the Relationship between High-Volume Hospitals and Outcomes in Surgery: A Systematic Review of Systematic Reviews. *Journal of the American College of Surgeons* 216(5): 1015-1025, 1025.e1-1025.e18. DOI: 10.1016/j.jamcollsurg.2012.12.049.

R Core Team (2019): R: A Language and Environment for Statistical Computing [*Open Source Software*]. R version 3.6.1 (Action of the Toes). Vienna: R Foundation for Statistical Computing. URL: https://cran.r-project.org/ [Download R for Windows > base > Download R 3.6.1 for Windows] (abgerufen am: 02.12.2019).

Schräder, P; Grouven, U; Bender, R (2007): Können Mindestmengen für Knieprothesen anhand von Routinedaten errechnet werden? Ergebnisse einer Schwellenwertanalyse mit Daten der externen stationären Qualitätssicherung. *Der Orthopäde* 36(6): 570-576. DOI: 10.1007/s00132-007-1066-7.

Stefoski Mikeljevic, J; Haward, RA; Johnston, C; Sainsbury, R; Forman, D (2003): Surgeon workload and survival from breast cancer. *British Journal of Cancer* 89(3): 487-491. DOI: 10.1038/sj.bjc.6601148.

Stroup, WW (2016): Generalized Linear Mixed Models: Modern Concepts, Methods and Applications. Boca Raton, US-FL [u. a.]: CRC Press. ISBN: 978-1-43-981513-7.

Taylor, HD; Dennis, DA; Crane, HS (1997): Relationship Between Mortality Rates and Hospital Patient Volume for Medicare Patients Undergoing Major Orthopaedic Surgery of the Hip, Knee, Spine, and Femur. *The Journal of Arthroplasty* 12(3): 235-242. DOI: 10.1016/S0883-5403(97)90018-8.

Townsend, Z; Buckley, J; Harada, M; Scott, MA (2013): The Choice Between Fixed and Random Effects. Chapter 5. In: Scott, MA; Simonoff, JS; Marx, BD; Hrsg.: *The SAGE Handbook of Multilevel Modeling.* London [u. a.]: SAGE Publications, 73-88. ISBN: 978-0-85702-564-7. DOI: 10.4135/9781446247600.

Urbach, DR; Austin, PC (2005): Conventional models overestimate the statistical significance of volume-outcome associations, compared with multilevel models. *Journal of Clinical Epidemiology* 58(4): 391-400. DOI: 10.1016/j.jclinepi.2004.12.001.

Varagunam, M; Hutchings, A; Black, N (2015): Relationship Between Patient-reported Outcomes of Elective Surgery and Hospital and Consultant Volume. *Medical Care* 53(4): 310-316. DOI: 10.1097/MLR.0000000000000318.

Wetzel, H (2006): Mindestmengen zur Qualitätssicherung: Konzeptionelle und methodische Überlegungen zur Festlegung und Evaluation von Fallzahlgrenzwerten für die klinische Versorgung. *Zeitschrift für Ärztliche Fortbildung und Qualität im Gesundheitswesen* 100(2): 99-106.

Wood, S (2019): mgcv: Mixed GAM Computation Vehicle with Automatic Smoothness Estimation [*Open Source Software*]. R package version 1.8-31. Published: 09.11.2019. Vienna: R Foundation for Statistical Computing. URL: https://CRAN.R-project.org/package=mgcv (abgerufen am: 02.12.2019).

Wood, SN (2006): Generalized Additive Models. An Introduction with R. (Texts in Statistical Science). Boca Raton, US-FL [u. a.]: Chapman & Hall/CRC. ISBN: 978-1-58488-474-3.

Wood, SN (2012): On *p*-values for smooth components of an extended generalized additive model. *Biometrika* 100(1): 221-228. DOI: 10.1093/biomet/ass048.

Wood, SN; Pya, N; Säfken, B (2016): Smoothing Parameter and Model Selection for General Smooth Models. *Journal of the American Statistical Association* 111(516): 1548-1563. DOI: 10.1080/01621459.2016.1180986.

Wood, SN (2017): Generalized Additive Models. An Introduction with R. Second Edition. (Texts in Statistical Science). Boca Raton, US-FL [u. a.]: Chapman & Hall/CRC. ISBN: 978-1-4987-2833-1.