Institut für Qualitätssicherung und
Transparenz im Gesundheitswesen

Weiterentwicklung des Strukturierten Dialogs mit Krankenhäusern

Abschlussbericht zu Stufe 1 und Stufe 2

Erstellt im Auftrag des
Gemeinsamen Bundesausschusses

Stand: 11. Februar 2020

Impressum

Thema:
Weiterentwicklung des Strukturierten Dialogs mit Krankenhäusern. Abschlussbericht zu Stufe 1 und Stufe 2

Ansprechpartnerin:
Andrea Wolf

Auftraggeber:
Gemeinsamer Bundesausschuss

Datum des Auftrags:
18. Januar 2018

Datum der Abgabe:
31. Januar 2020; redaktionell geändert am 11. Feburar 2020

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

Abkürzungsverzeichnis

Abkürzung	Bedeutung
aQua-Institut	Institut für angewandte Qualitätsförderung und Forschung im Gesundheitswesen GmbH
BQB	Bundesqualitätsbericht
DeQS-RL	Richtlinie zur datengestützten einrichtungsübergreifenden Qualitätssicherung
EJ	Erfassungsjahr
EJ + 1	Jahr, das ein Jahr nach dem Erfassungsjahr endet (usw. EJ + 2, EJ + 3)
esQS	externe stationäre Qualitätssicherung
G-BA	Gemeinsamer Bundesausschuss
GKV-FQWG	GKV-Finanzstruktur- und Qualitätsweiterentwicklungsgesetzt (Gesetz zur Weiterentwicklung der Finanzstruktur und der Qualität in der gesetzlichen Krankenversicherung)
ID	Identifikationsnummer
IQTIG	Institut für Qualitätssicherung und Transparenz im Gesundheitswesen
KHSG	Krankenhausstrukturgesetz (Gesetz zur Reform der Strukturen der Krankenhausversorgung)
LAG	Landesarbeitsgemeinschaft(en) für sektorenübergreifende Qualitätssicherung
LE	Leistungserbringer
LQS	Landesgeschäftsstelle(n) für Qualitätssicherung
O/E	Verhältnis aus der beobachteten und der erwarteten Rate (*observed to expected ratio*)
plan. QI-RL	Richtlinie zu planungsrelevanten Qualitätsindikatoren
PPV	Positive-Predictive-Value
QB	Qualitätsbewertung
Qesü-RL	Richtlinie zur einrichtungs- und sektorenübergreifenden Qualitätssicherung
QFD-RL	Qualitätsförderungs- und Durchsetzungs-Richtlinie
QI	Qualitätsindikator
QIDB	Qualitätsindikatorendatenbank
QS	Qualitätssicherung

Abkürzung	Bedeutung
QS-Verfahren	Qualitätssicherungsverfahren
QS CHE	QS-Verfahren *Cholezystektomie*
QS PCI	QS-Verfahren *Perkutane Koronarintervention (PCI) und Koronarangiographie*
QS WI	QS-Verfahren *Vermeidung nosokomialer Infektionen – postoperative Wundinfektionen*
QSEB	Qualitätssicherungsergebnisberichte
QSKH-RL	Richtlinie über Maßnahmen der Qualitätssicherung in Krankenhäusern
RL	Richtlinie
SD	Strukturierter Dialog
SGB V	Fünftes Buch Sozialgesetzbuch
SIR	standardisiertes Inzidenzverhältnis
SMR	standardisiertes Morbititätsverhältnis
SOP	Standardvorgehensweise (*Standard Operating Procedure*)

1 Einleitung

Mit Beschluss vom 18. Januar 2018 hat der Gemeinsame Bundesausschuss (G-BA) das Institut für Qualitätssicherung und Transparenz im Gesundheitswesen (IQTIG) mit der Weiterentwicklung des Strukturierten Dialogs mit Krankenhäusern beauftragt. Dieser Weiterentwicklungsauftrag ist in zwei Stufen mit unterschiedlichen Inhalten und Abgabefristen gegliedert. Der Abschlussbericht zu Stufe 1 wurde dem G-BA am 30. November 2018 übermittelt. Der vorliegende Bericht beinhaltet sowohl die weiterentwickelten Empfehlungen des Berichts zu Stufe 1 als auch die für Stufe 2 beauftragten Inhalte. Aus Gründen der Lesbarkeit wurde davon abgesehen, Veränderungen im Vergleich zu dem Bericht zu Stufe 1 hervorzuheben.

Im Folgenden werden zunächst der Hintergrund der Beauftragung, Schnittstellen zu anderen Beauftragungen sowie das Auftragsverständnis des IQTIG für beide Stufen der Beauftragung dargestellt. Anschließend wird auf die Einbeziehung der auf Landesebene zuständigen LAG- und LQS-Vertreterinnen und -Vertreter und weiterer externer Expertise sowie den Ablauf des Stellungnahmeverfahrens zum Abschlussbericht eingegangen.

1.1 Hintergrund

Als Strukturierter Dialog werden diejenigen Prozesse bezeichnet, die gemäß §§ 11 bis 14 QSKH-RL eingeleitet werden sollen, wenn ein Leistungserbringer in einem Qualitätsindikator oder einem Auffälligkeitskriterium der QSKH-RL vom Referenzwert abweicht. Laut § 12 QSKH-RL soll dadurch geprüft werden, ob die dokumentierten Leistungen in der fachlich gebotenen Qualität erbracht und korrekt dokumentiert wurden.

Die Verantwortung für die Durchführung des Strukturierten Dialogs liegt bei der jeweils mit der Durchführung des Verfahrens beauftragten Stelle. Dies ist für einige Qualitätssicherungsverfahren (QS-Verfahren) der Unterausschuss Qualitätssicherung des G-BA (sog. direkte bzw. bundesbezogene Verfahren) und für den Großteil der QS-Verfahren (sog. indirekte bzw. landesbezogene Verfahren) das verantwortliche Gremium der Landesgeschäftsstellen für Qualitätssicherung (LQS) bzw. der Landesarbeitsgemeinschaften (LAG) in den Bundesländern.

Analysen der Ergebnisse des Strukturierten Dialogs, die das IQTIG dem G-BA jährlich im Mai vorlegt, deuten auf eine deutliche Heterogenität im Vorgehen und in den Ergebnissen zwischen den beauftragten Stellen hin (IQTIG 2018a; siehe auch Abschnitt 2.5). Faire Leistungserbringervergleiche und aussagekräftige Ergebnisse erfordern jedoch ein einheitliches, standardisiertes Vorgehen aller beauftragten Stellen. Gleichzeitig wächst der Aufwand der beauftragten Stellen mit steigender Anzahl von QS-Verfahren und Indikatoren. Um die Prozesse des Strukturierten Dialogs stärker zu standardisieren und den Aufwand zu reduzieren wurden 2013 in einem gemeinsamen Workshop des G-BA mit den LQS Schwachstellen und Verbesserungspotenziale identifiziert. Dieser Weiterentwicklungsbedarf wurde im Auftrag des G-BA im Rahmen von drei Projektgruppen weiter konkretisiert. Die Ergebnisse dieser Projektgruppen mündeten 2014 in

Vorschläge für Aufträge für die damalige Institution nach § 137a SGB V[1]. Eine Beauftragung erfolgte jedoch aufgrund der Beendigung des Vertrags mit dem AQUA-Institut damals nicht.

1.2 Schnittstellen zu anderen Beauftragungen und Richtlinien

Am 19. Juli 2018 hat der G-BA die Richtlinie zur datengestützten einrichtungsübergreifenden Qualitätssicherung (DeQS-RL)[2] beschlossen (G-BA 2018b), in die die indikatorbasierten QS-Verfahren sowohl der bestehenden Richtlinie zur sektorenübergreifenden Qualitätssicherung (Qesü-RL)[3] als auch der sektorenspezifischen Richtlinie über Maßnahmen der Qualitätssicherung in Krankenhäusern (QSKH-RL) überführt werden sollen.

Seit dem 1. Januar 2019 ist die DeQS-RL in Kraft und hat die Qesü-RL abgelöst. Damit sind jetzt die bisher in der Qesü-RL verorteten Indikatorensets den Regelungen der DeQS-RL unterworfen. Gleiches wird zukünftig sukzessive für die QS-Verfahren im Rahmen der QSKH-RL erfolgen. Das bedeutet, dass in der Übergangszeit einige QS-Verfahren noch nach den Vorgaben der QSKH-RL und einige QS-Verfahren bereits nach den Vorgaben der DeQS-RL durchgeführt werden. Darüber hinaus werden seit 2017 einige Qualitätsindikatoren der QSKH-RL auch nach den Regelungen der Richtlinie für planungsrelevante Qualitätsindikatoren (plan. QI-RL) verwendet, um den für die Krankenhausplanung zuständigen Landesbehörden qualitätsorientierte Entscheidungen zu ermöglichen. Gemeinsam ist allen vier Richtlinien, dass Qualitätsindikatoren verwendet werden, um leistungserbringerbezogene Aussagen zur Versorgungsqualität in bestimmten Themenbereichen zu machen.

Mit Beschluss vom 17. Mai 2018 wurde das IQTIG durch den G-BA auch mit der „Entwicklung von Kriterien für die Datenbewertung und die Einleitung und Durchführung von Qualitätssicherungsmaßnahmen“ im Rahmen der Qesü-RL beauftragt. Dieser Auftrag beinhaltete Empfehlungen zu entwickeln für bundesweit einheitliche Bewertungskriterien für die Ergebnisse der Qualitätsindikatoren aus den QS-Verfahren *Perkutane Koronarintervention (PCI) und Koronarangiographie* (*QS PCI*) und *Vermeidung nosokomialer Infektionen – postoperative Wundinfektionen* (*QS WI*). In § 17 Qesü-RL ist geregelt, dass bei Auffälligkeiten in den Indikatorergebnissen durch die mit der Durchführung des Verfahrens zuständigen Stelle ein Stellungnahmeverfahren durchzuführen ist. Dieses Stellungnahmeverfahren wird im vorliegenden Bericht als Pendant zum Strukturierten Dialog der QSKH-RL verstanden. In der Qesü-RL ist jedoch weder definiert, was eine Auffälligkeit in einem Indikatorergebnis ist, noch ist der Ablauf des

[1] AQUA – Institut für angewandte Qualitätsförderung und Forschung im Gesundheitswesen GmbH; bis 31. Dezember 2015 die zuständige Institution nach § 137a SGB V.

[2] Richtlinie des Gemeinsamen Bundesausschusses gemäß § 136 Abs. 1 SGB V i. V. m. § 135a SGB V über Maßnahmen der Qualitätssicherung für nach § 108 SGB V zugelassene Krankenhäuser (Richtlinie über Maßnahmen der Qualitätssicherung in Krankenhäusern / QSKH-RL). In der Fassung vom 15. August 2006, zuletzt geändert am 16. März 2018, in Kraft getreten am 27. April 2018. URL: https://www.g-ba.de/informationen/richtlinien/38/ (abgerufen am: 05.07.2018).

[3] Richtlinie des Gemeinsamen Bundesausschusses nach § 92 Abs. 1 Satz 2 Nr. 13 i. V. m. § 136 Abs. 1 Nr. 1 SGB V über die einrichtungs- und sektorenübergreifenden Maßnahmen der Qualitätssicherung (Richtlinie zur einrichtungs- und sektorenübergreifenden Qualitätssicherung – Qesü-RL). In der Fassung vom 19. April 2010, zuletzt geändert am 19. Oktober 2017, in Kraft getreten am: 8. Februar 2018. URL: https://www.g-ba.de/informationen/richtlinien/72/ (abgerufen am: 28.11.2018).

Stellungnahmeverfahrens näher beschrieben. Die Beauftragung vom 17. Mai 2018 beinhaltet also eine Ausgestaltung des Stellungnahmeverfahrens nach § 17 Qesü-RL.

Um ein einheitliches Vorgehen bei der Auswertung und Bewertung von Indikatorergebnissen über verschiedene Richtlinien hinweg sicherzustellen, wurden die Empfehlungen für das Stellungnahmeverfahren nach § 17 Qesü-RL aus dem Bericht zu Stufe 1 der Beauftragung zur Weiterentwicklung des Strukturierten Dialogs mit Krankenhäusern abgeleitet. Zu dieser Beauftragung hat das IQTIG am 15. März 2019 den Abschlussbericht an den G-BA übermittelt. Die Qesü-RL trat jedoch schon mit Wirkung zum 1. Januar 2019 außer Kraft.

Die Beauftragung des IQTIG zur Weiterentwicklung des Strukturierten Dialogs mit Krankenhäusern beinhaltet die Entwicklung eines Rahmenkonzepts, das mit anderen relevanten Richtlinien abzustimmen ist; dies beinhaltet aus Sicht des IQTIG die DeQS-RL sowie die plan. QI-RL.[4] Da die Qesü-RL zum Zeitpunkt der Berichtlegung nicht mehr in Kraft ist und die QSKH-RL wie oben beschrieben auch mittelfristig außer Kraft gesetzt werden soll, erscheint es aus Sicht des IQTIG sinnvoll, ein einheitliches Rahmenkonzept für ein Stellungnahmeverfahren zu den quantitativen Ergebnissen der datengestützten Qualitätssicherung zu entwickeln, das prinzipiell in allen indikatorbasierten QS-Verfahren zur Anwendung kommen kann. Dies ist auch im Einklang mit der Beauftragung, die vorsieht, dass „Elemente des Strukturierten Dialogs […] zukünftig auch in der Rahmenrichtlinie zur datengestützten Qualitätssicherung [DeQS-RL] und in der Richtlinie zur qualitätsabhängigen Vergütung eingesetzt werden können [sollen].“ Mit diesem Ziel legt das IQTIG daher in diesem Abschlussbericht seine Empfehlungen für die Gestaltung eines Stellungnahmeverfahrens für die DeQS-RL[5] vor. Er beinhaltet konkrete Empfehlungen für die Auswertung und Bewertung von leistungserbringerbezogenen Qualitätsindikatorergebnissen auf Basis von Dokumentationsdaten der Leistungserbringer, auf Basis von Sozialdaten sowie auf Basis durch Patientenbefragungen[6] erhobener Daten.

1.3 Auftragsverständnis

Das IQTIG wurde beauftragt, Empfehlungen für die Weiterentwicklung des Strukturierten Dialogs mit Krankenhäusern zu entwickeln. Laut Auftrag verfolgt der G-BA damit die folgenden drei Ziele: Reduktion der Heterogenität in der Vorgehensweise zwischen den beauftragten Stellen, Erhöhung der Transparenz und Nachvollziehbarkeit der Entscheidungen sowie Steigerung der Effizienz des Strukturierten Dialogs.

Unter „1. Optimierung der Einheitlichkeit der Vorgehensweise“ ist das IQTIG beauftragt, einheitliche Kriterien für die Einholung von Stellungnahmen sowie Empfehlungen zu Form und Inhalt der Stellungnahmen zu entwickeln. Darüber hinaus sollen Kriterien erarbeitet werden, die die Bewertung von Stellungnahmen vereinheitlichen und erleichtern sollen. In diesem Kontext soll

[4] Eine Abstimmung mit einer „Richtlinie zur qualitätsabhängigen Vergütung“ ist auch gefordert. Diese lag zum Zeitpunkt der Berichtserstellung allerdings noch nicht vor.

[5] Aus diesem Grund werden ab Kapitel 4 „Eckpunkte des Rahmenkonzepts für die Qualitätsbewertung und -förderung“ nicht mehr die Begriffe der QSKH-RL sondern die der DeQS-RL verwendet und der Begriff Strukturierter Dialog durch Stellungnahmeverfahren ersetzt.

[6] Die Auswertungsmethodik für Qualitätsindikatoren auf Basis von Patientenbefragungsdaten ist in den jeweiligen Abschlussberichten beschrieben.

auch das bestehende „Schema zur Einstufung und Bewertung von Indikatorergebnissen“ überarbeitet werden.

In diesem Zusammenhang sollen auch Empfehlungen für die Anforderungen an die Fachexpertinnen und Fachexperten sowie die Zusammensetzung der Expertengremien als auch Möglichkeiten, um die Auswirkung von Interessenkonflikten der Fachexpertinnen und Fachexperten zu minimieren, entwickelt werden.

Die Empfehlungen für „1. Optimierung der Einheitlichkeit der Vorgehensweise“ finden sich in Abschnitt 5.6 und Kapitel 6.

Das IQTIG ist weiterhin beauftragt, Empfehlungen zur Evaluierung des erarbeiteten Konzepts zu geben. Obwohl unter 1. aufgeführt, wird dies als Auftrag verstanden, Evaluationsempfehlungen für das gesamte Konzept zu erarbeiten. Die Empfehlungen zur Evaluation finden sich in Kapitel 10.

Unter „2. Optimierung der Transparenz und Nachvollziehbarkeit der Entscheidungsfindung“ soll die Berichterstattung zu den Ergebnissen des Strukturierten Dialogs weiterentwickelt werden. Dies beinhaltet gemäß der Beauftragung Vorschläge, wie die gleiche Datengrundlage für den Qualitätsreport und für den Bericht zum Strukturierten Dialog, den das IQTIG regelmäßig im Mai auf Basis der Berichte der Stellen auf Landesebene erstellt, gewährleistet werden kann. Zusätzlich sollen Mindestanforderungen für Berichte der Landesstellen formuliert werden. Während im Bericht zu Stufe 1 der Beauftragung noch Empfehlungen für die Berichte der QSKH-RL gemacht wurden, wird in dem vorliegenden Abschlussbericht davon abgesehen, diese Empfehlungen weiterzuentwickeln. Vor dem Hintergrund des absehbaren Außerkrafttretens der QSKH-RL erscheint der Aufwand, der eine Umgestaltung der Berichterstattung zu den Qualitätsergebnissen im Rahmen dieser Richtlinie bedeutet, nicht mehr gerechtfertigt. Daher werden auch an dieser Stelle die Empfehlungen mit Blick auf die DeQS-RL gegeben. Die Empfehlungen zu „2. Optimierung der Transparenz und Nachvollziehbarkeit der Entscheidungsfindung“ finden sich in Kapitel 9.

Unter 3. „Optimierung der Effizienz des Verfahrens“ wird die Optimierung der „Diskriminationsfähigkeit“ der Indikatoren unter Prüfung verschiedener statistischer Verfahren beauftragt. Unter Diskriminationsfähigkeit versteht das IQTIG in diesem Kontext die Güte, mit der Leistungserbringer auf Basis eines Indikatorergebnisses in solche mit Qualitätsdefizit im Sinne des Qualitätsziels des jeweiligen Indikators und solche Leistungserbringer ohne Qualitätsdefizit unterteilt werden können. Aus Sicht des IQTIG spielen allerdings neben dem statistischen Verfahren auch andere Faktoren eine Rolle für die Effizienz des Verfahrens: Da mit zunehmender Standardisierung von Prozessen typischerweise der für die Durchführung benötigte Aufwand sinkt, versteht das IQTIG die zur Vereinheitlichung der Vorgehensweise zu entwickelnden Empfehlungen als weitere Faktoren zur Steigerung der Effizienz. Die Empfehlungen zur Optimierung der Effizienz finden sich dementsprechend im Abschnitt 5.6 und Kapitel 6.

Außerdem sollen Möglichkeiten zur Verkürzung des Verfahrens geprüft werden. Hierunter versteht das IQTIG Möglichkeiten zur Verkürzung des Stellungnahmeverfahrens als Teil des Gesamtverfahrens. Möglichkeiten zur Verkürzung des Stellungnahmeverfahrens finden sich in Kapitel 8.

Unter „3. Optimierung der Effizienz des Verfahrens“ sind außerdem die Prüfung der Verwendbarkeit von Indizes als Aufgreifkriterium für die Qualitätsbewertung von Leistungserbringern innerhalb eines Leistungsbereichs sowie Optionen für leistungsbereichübergreifende Indikatorensets beauftragt. Unter Indizes versteht das IQTIG das Ergebnis einer gewichteten Aggregation von Variablen, in diesem Fall Qualitätsindikatoren. Dazu erfolgt in Abschnitt 2.2 eine Diskussion des Verständnisses von Qualitätsindikatoren als Aufgreifkriterien gegenüber Qualitätsindikatoren als statistische Kenngrößen für Qualität. Die Empfehlungen zu diesen Punkten der Beauftragung finden sich in den Abschnitten 11.3 und 11.4.

Darüber hinaus wurde die methodische Weiterentwicklung aller Indikatoren mit beauftragt. Erste Empfehlungen, wie eine solche Weiterentwicklung durchgeführt werden sollte, finden sich in Abschnitt 11.1.

Unter Punkt 4. des Auftrags wird die Entwicklung eines Rahmenkonzepts beauftragt, das die Elemente des Strukturierten Dialogs weiterentwickelt und mit anderen relevanten Richtlinien abstimmt. Das IQTIG versteht diesen Punkt der Beauftragung als den konzeptuellen Rahmen, in dem auf Basis des Grundverständnisses von Qualitätsmessungen und deren Gütekriterien die zentralen Festlegungen zusammengefasst werden, aus denen sich die konkreten Empfehlungen ableiten lassen. Wie in Abschnitt 1.2 dargestellt, wurden aufgrund des Auslaufens der QSKH-RL und der außer Kraft gesetzten Qesü-RL ein richtlinienunabhängiges Rahmenkonzept und Empfehlungen entwickelt, die sich an den Begrifflichkeiten der DeQS-RL orientieren und damit auch die vertragsärztliche Versorgung betreffen. Die Eckpunkte dieses Rahmenkonzepts finden sich in Kapitel 4.

Die sich an die Qualitätsbewertung anschließenden Maßnahmen zur Qualitätsverbesserung sind nach Verständnis des IQTIG nicht Teil der Beauftragung. Weder in den Zielen der Beauftragung noch in den Beauftragungspunkten 1 bis 4 wird eine Weiterentwicklung der Qualitätsförderungskomponente des Strukturierten Dialogs angesprochen. Da die Steigerung der Versorgungsqualität allerdings ein zentrales Ziel der gesetzlichen Qualitätssicherung darstellt, legt das IQTIG auch erste Empfehlungen für das Einleiten von qualitätsfördernden Maßnahmen vor. Dies beinhaltet Empfehlungen für Kriterien zum Abschluss und zum Inhalt von Zielvereinbarungen sowie Instrumente zur mittel- bis längerfristigen Überprüfung der Zielerreichung. Da qualitätsfördernde Maßnahmen an die identifizierten Verbesserungspotenzialen der betroffenen Leistungserbringer individuell angepasst werden sollten, sind diese jedoch schwieriger zu standardisieren als Qualitätsmessungen. Da individuell gestaltete qualitätsverbessernde Maßnahmen erfahrungsgemäß zeit- und ressourcenintensiv sind, versteht das IQTIG den Auftrag auch dahingehend, dass durch eine Verschlankung des Vorgehens bei der Qualitätsbewertung mehr Ressourcen für anschließende Maßnahmen der Qualitätsverbesserung zur Verfügung stehen können. Diese Empfehlungen für qualitätsverbessernde Maßnahmen finden sich in Kapitel 7.

Obwohl nach § 9 Abs. 7 QSKH-RL auch bei Auffälligkeiten in der Datenvalidierung der Strukturierte Dialog eingeleitet werden soll, sieht die Beauftragung nicht explizit das Erarbeiten eines Konzepts zur Bewertung der Daten- und Dokumentationsqualität vor. Nach Verständnis des IQTIG ist dies damit nicht Teil des Auftrags. Die Weiterentwicklung des Vorgehens bei Auffälligkeiten in der Datenvalidierung im Rahmen der DeQS-RL wird jedoch empfohlen.

1.4 Einbeziehung der auf Landesebene zuständigen LAG- und LQS-Vertreterinnen und -Vertreter

Die Beauftragung verlangt auch, die auf Landesebene mit der Durchführung des Strukturierten Dialogs nach QSKH-RL beauftragten Stellen (LQS) und soweit wie möglich die Landesarbeitsgemeinschaften (LAG) in die Weiterentwicklung einzubeziehen. Dies wurde durch einen Workshop mit den Vertreterinnen und Vertretern der oben genannten Stellen umgesetzt[7].

Der Workshop fand am 27. Mai 2019 in den Räumlichkeiten des IQTIG statt. Ziel des Workshops war es, die Eckpunkte des Konzepts vorzustellen und die Expertise der zuständigen Teilnehmerinnen und Teilnehmer in der Durchführung des Strukturierten Dialogs sowie deren Rückmeldungen zum Konzept einzuholen, um diese in die weitere Bearbeitung von Stufe 2 der Beauftragung einfließen zu lassen. Der Workshop wurde sieben Wochen vor Beginn allen LQS- und LAG-Vertreterinnen und -Vertretern angekündigt. Insgesamt nahmen 19 LAG- und LQS-Vertreterinnen und -vertreter aus 14 Bundesländern sowie die beim IQTIG für die direkten Verfahren zuständigen Personen am Workshop teil. Zur Vorbereitung auf den Workshop wurden allen Teilnehmerinnen und Teilnehmer zwei Wochen vor dem Workshop eine schriftliche Zusammenfassung des Rahmenkonzepts und die Tagesordnung zugesandt.

Im Workshop präsentierte das IQTIG das Rahmenkonzept und diskutierte mit den Teilnehmerinnen und Teilnehmern mögliche Herausforderungen sowie Optionen zur Verkürzung des Verfahrens. Der Verlauf und die Ergebnisse des Workshops wurden in einem Protokoll festgehalten und die Ergebnisse für die Weiterentwicklung berücksichtigt (siehe Anhang, Kapitel 1).

Das IQTIG präsentierte das Rahmenkonzept und die Ergebnisse des Workshops unter Beteiligung der LAG-Vertreterinnen und -Vertreter auch im darauffolgenden LQS-IQTIG-Treffen am 18. Juni 2019 (siehe Anhang, Kapitel 2).

1.5 Einbeziehung weiterer externer Expertise

Gemäß der in den Methodischen Grundlagen V1.1 festgehaltenen Vorgehensweise zur „Einbeziehung externen Wissens und externer Erfahrung“ wurde zusätzliche externe Expertise für die Entwicklung der statistischen Methoden (Kapitel 5) herangezogen. Als Fachexperte für statistische Prozesskontrolle wurde Prof. Dr. Sven Knoth vom Lehrstuhl für Rechnergestützte Statistik der Helmut-Schmidt-Universität, Hamburg, konsultiert. Im Rahmen eines eintägigen Workshops

[7] In den unterschiedlichen Kapiteln wird auf die seitens der Ländervertreterinnen und -vertreter genannten Herausforderungen (Ergebnisse des Workshops) Bezug genommen.

wurden zusammen mit Prof. Dr. Sven Knoth die unter Punkt 3.a des Auftrags genannten Fragestellungen diskutiert und statistische Lösungsansätze evaluiert. Als Ergebnis dieser Beratung stehen die Empfehlungen in Abschnitt 5.6.

1.6 Stellungnahmeverfahren zum Vorbericht

Gemäß § 137a Abs. 7 SGB V sind bei Entwicklungsarbeiten des IQTIG verschiedene Organisationen, Verbände und Interessenvertretungen des Gesundheitswesens zu beteiligen. Um dieser gesetzlichen Pflicht nachzukommen, führte das IQTIG ein Stellungnahmeverfahren für den gemeinsamen Abschlussbericht zu Stufe 1 und Stufe 2 der Beauftragung durch. Die folgenden Organisationen, Verbände und Interessenvertretungen erhielten vom 10. Oktober 2019 bis zum 21. November 2019 die Gelegenheit, schriftlich Stellung zu den Inhalten dieses Berichts zu nehmen:

- die Kassenärztliche Bundesvereinigung
- die Deutsche Krankenhausgesellschaft
- der Spitzenverband Bund der Krankenkassen
- der Verband der Privaten Krankenversicherung
- die Bundesärztekammer
- die Bundeszahnärztekammer
- die Bundespsychotherapeutenkammer
- die Berufsorganisationen der Krankenpflegeberufe
- die wissenschaftlichen medizinischen Fachgesellschaften
- das Deutsche Netzwerk Versorgungsforschung
- die für die Wahrnehmung der Interessen der Patientinnen und Patienten und der Selbsthilfe chronisch kranker und behinderter Menschen maßgeblichen Organisationen auf Bundesebene
- der oder die Beauftragte der Bundesregierung für die Belange der Patientinnen und Patienten
- zwei von der Gesundheitsministerkonferenz der Länder zu bestimmende Vertreter
- die Bundesoberbehörden im Geschäftsbereich des Bundesministeriums für Gesundheit, soweit ihre Aufgabenbereiche berührt sind
- die mit der Durchführung der QS-Verfahren gemäß QSKH-RL sowie DeQS-RL beauftragten Stellen auf Landesebene

Die eingegangenen Stellungnahmen werden vom IQTIG schriftlich gewürdigt und der Bericht wurde auf Basis der Stellungnahmen auf Anpassungsbedarf hin geprüft. Der vorliegende Abschlussbericht beinhaltet alle Anpassungen, die sich aus den Stellungnahmen zu dem Vorbericht ergeben haben.

2 Analyse der Ausgangssituation

Um die Weiterentwicklung des Strukturierten Dialogs an den Zielen der Qualitätssicherung auszurichten, wird im vorliegenden Kapitel eine Ist-Analyse vorgenommen, in der die aktuelle Vorgehensweise gemäß QSKH-RL beschrieben und methodisch eingeordnet wird. Obwohl sich die Empfehlungen in den folgenden Kapiteln auf ein richtlinienunabhängiges Stellungnahmeverfahren beziehen (vgl. Abschnitt 1.3) und diese aufgrund des Auslaufens der QSKH-RL vor allem mit Blick auf das in § 17 DeQS-RL beschriebene Stellungnahmeverfahren entwickelt wurden, wird eine Analyse des Strukturierten Dialogs gemäß QSKH-RL als sinnvoll erachtet. Einerseits finden sich die in Abschnitt 2.1 identifizierte Funktionen des Strukturierten Dialogs auch in ähnlicher Form in dem Stellungnahmeverfahren der DeQS-RL. Andererseits lässt § 17 DeQS-RL ähnlich wie die QSKH-RL Spielräume in der Ausgestaltung dieses Stellungnahmeverfahrens, welche potenzielle Quellen für Heterogenität in der Umsetzung der Richtlinie darstellen.

2.1 Funktionen des Strukturierten Dialogs

In den §§ 11 bis 14 QSKH-RL sind die Regelungen zum Strukturierten Dialog mit Krankenhäusern festgehalten. Allerdings beinhaltet die QSKH-RL keine Definition des Begriffs „Strukturierter Dialog“. Im Beschluss zur Beauftragung des IQTIG wird der Strukturierten Dialog beschrieben als Verfahren, „in dem unter Einbeziehung von Experten auffällige Ergebnisse von Einrichtungen bewertet und erforderlichenfalls geeignete Maßnahmen zur Qualitätsverbesserung und -förderung eingeleitet werden“ (G-BA 2018a). Zudem sieht die QSKH-RL vor, dass im Rahmen des Strukturierten Dialogs auch eine Überprüfung der Validität der Dokumentation erfolgt (§ 12 Abs. 1 QSKH-RL). Aus den §§ 11 bis 14 sowie § 9 Abs. 7 QSKH-RL lassen sich für den Begriff „Strukturierter Dialog“ mehrere Prozesse mit unterschiedlichen Funktionen und Zielen ableiten und zusammenfassen:

- **Bewertung der Versorgungsqualität**
  Der Strukturierte Dialog dient bisher dazu, anhand der Ergebnisse von Qualitätsindikatoren und schriftlichen Stellungnahmen der Leistungserbringer sowie aufgrund von Begehungen und kollegialen Gesprächen die Versorgungsqualität der Leistungserbringer zu bewerten.
- **Bewertung der Dokumentationsqualität**
  Außerdem erfolgt eine Beurteilung der Dokumentationsqualität im Rahmen des Datenvalidierungsverfahrens und der Prüfung der schriftlichen Stellungnahmen hinsichtlich angegebener Dokumentationsfehler. Der Strukturierte Dialog erfüllt demnach auch die Funktion der Bewertung der Dokumentationsqualität der einzelnen Leistungserbringer.
- **Qualitätsförderung**
  Die Steigerung der Qualität von Krankenhausleistungen ist darüber hinaus ebenso ein Ziel des Strukturierten Dialogs. Zu diesem Zweck kennt die QSKH-RL Maßnahmen, die zur Steigerung der Versorgungsqualität eingeleitet werden können, wie etwa Zielvereinbarungen.

Sensitivität und Spezifität

Neben diesen explizit in der QSKH-RL festgehaltenen Funktionen hat der Strukturierte Dialog auch die implizite Funktion, die Spezifität der Qualitätsmessung zu steigern. Zur Illustration wird diese Funktion anhand eines Qualitätsindikators und unter folgenden Annahmen erläutert:

- Jeder Leistungserbringer weist bezogen auf das Qualitätsziel eines Indikators eine zugrundeliegende „tatsächliche“ Qualität auf.
- Das Vorgehen besteht aus einem ersten indikatorbasierten Prüfschritt und einem nachgelagerten zweiten Prüfschritt (Strukturierter Dialog).
- Das Gesamtverfahren bestehend aus Prüfschritt 1 und 2 attestiert am Ende entweder unauffällige oder auffällige Qualität bezogen auf das Qualitätsziel des betreffenden Indikators.

Daraus ergeben sich am Ende des Gesamtverfahrens die in Tabelle 1 dargestellten Möglichkeiten von richtigen und falschen Klassifikationen. Als Sensitivität des Vorgehens wird dabei der Anteil der mit „auffällige Qualität“ bewerteten Leistungserbringerergebnisse an allen Leistungserbringerergebnissen mit „tatsächlich“ unzureichender Qualität bezeichnet. Als Spezifität des Vorgehens wird der Anteil der mit „unauffällige Qualität“ bewerteten Leistungserbringerergebnissen an allen Leistungserbringerergebnissen mit „tatsächlich“ zureichender Qualität bezeichnet. Diese Betrachtung ist allerdings hypothetisch, da einerseits kein Goldstandard für die Detektion „tatsächlicher“ Qualitätsdefizite vorliegt und andererseits eine Definition „tatsächlicher“ Qualität nur schwer möglich ist.[8]

Tabelle 1: Sensitivität und Spezifität des Gesamtverfahrens zur Qualitätsmessung bezogen auf einen Qualitätsindikator

		„tatsächliche“ Qualität	
		unzureichend	zureichend
Qualitätsbewertung	auffällig	richtig positiv	falsch positiv
	unauffällig	falsch negativ	richtig negativ

Ein auffälliges Indikatorergebnis (Ergebnis erster Prüfschritt) kann im Rahmen des Strukturierten Dialogs (zweiter Prüfschritt) zwar widerlegt werden, bei einem unauffälligen Indikatorergebnis kann dieses jedoch nicht widerlegt werden, da nur auffällige Ergebnisse einem zweiten Prüfschritt unterzogen werden. In Tabelle 1 entspricht dies einer Verschiebung einiger Ergebnisse des ersten Prüfschritts von den „falsch positiven“ hin zu den „richtig negativen“

[8] Dahinter steht die Frage, ob es eine von der konkreten Messung unabhängige, „tatsächliche“ Qualität der Versorgung theoretisch geben kann, oder ob Qualität nur mit Bezug zu der jeweiligen Messung definiert werden kann.

durch den zweiten Prüfschritt. Daher wird durch den Strukturierten Dialog die Spezifität des Vorgehens erhöht, im Vergleich zu einem Vorgehen ohne Strukturierten Dialog. Die Sensitivität des Gesamtverfahrens wird dagegen maßgeblich durch den ersten Prüfschritt bestimmt.

Eine Steigerung der Spezifität durch einen zweiten Prüfschritt ohne Steigerung der Sensitivität ist dann sinnvoll, wenn der erste Prüfschritt eine hohe Sensitivität aber eine vergleichsweise niedrigere Spezifität aufweist. Bei der Konstruktion von Qualitätsindikatoren kann die Vielfältigkeit der Gesundheitsversorgung zu unterschiedlichem Ausmaß berücksichtigt werden. Je stärker diese Vielfältigkeit schon bei der Definition von Zähler, Nenner sowie Ein- und Ausschlusskriterien des Indikators berücksichtigt wird, desto höher ist die Spezifität des Qualitätsindikators im obigen Sinn.

Damit einher geht jedoch auch eine Steigerung in der Komplexität des Indikators, was häufig auch größeren Erhebungsaufwand bedeutet. Dies wird am Beispiel eines Qualitätsindikators „Thromboseprophylaxe“ illustriert, der den Anteil von Patientinnen und Patienten mit Thromboseprophylaxe an allen operierten Patientinnen und Patienten abbildet. Als Ausschlusskriterium könnte die Einnahme von Medikamenten zur Blutverdünnung entweder direkt miterfasst werden, oder erst im zweiten Prüfschritt für die dokumentierten Fälle überprüft werden. Wird dieses Ausschlusskriterium erst im zweiten Prüfschritt berücksichtigt, ist die Klassifikation von Leistungserbringern in diejenigen mit unauffälliger und diejenigen mit auffälliger Qualität im ersten indikatorbasierten Schritt weniger spezifisch im obigen Sinn als wenn dies bereits durch den Indikator berücksichtigt würde. Eine Steigerung der Spezifität des Vorgehens kann also aufgrund einer als zu niedrig empfundenen Spezifität der indikatorbasierten Klassifikation als erforderlich erachtet werden. Anhand dieses Beispiels wird auch deutlich, dass eine Aufwands- und Spezifitätsverschiebung vom ersten Prüfschritt hin zu dem zweiten erfolgt, sofern das Ausschlusskriterium erst im zweiten Prüfschritt berücksichtigt wird. Indikatoren, die wie in diesem Beispiel weniger „spezifisch“ sind, eignen sich damit eher als „Aufgreifkriterien der Versorgungsqualität“ für einen darauffolgenden Prüfschritt und weniger als abschließende Qualitätsmessungen. Denn wie im Folgenden erläutert wird, unterscheiden sich die methodischen Anforderungen an Qualitätsmessungen je nach Verwendungszweck (IQTIG 2019a, S. 38).

2.2 Anforderungen an die Qualitätsmessung und -bewertung

Die Anforderungen, denen die Prozesse und Ergebnisse der gesetzlich verpflichtenden Qualitätssicherung genügen müssen, haben sich von ihren Anfängen bis heute deutlich gesteigert. Während in den Anfängen der Qualitätssicherungsbemühungen im Gesundheitswesen das Lernen von den Besseren in einem geschützten Raum im Vordergrund stand, wurden durch verschiedene Gesetzgebungen mehr und mehr die Transparenz, öffentliche Berichterstattung und Steuerungsfunktion von Versorgungsqualität in den Vordergrund gerückt (siehe auch I. Hintergrund zur Beauftragung im Beschluss zur Beauftragung).

So wurde im Jahr 2014 mit dem GKV-Finanzstruktur- und Qualitäts-Weiterentwicklungsgesetz (GKV-FQWG) beschlossen, dass eine leistungserbringervergleichende Übersicht über die Qualität der stationären Versorgung im Internet veröffentlichen werden soll. Mit dem im darauffolgenden Jahr beschlossenen Krankenhausstrukturgesetz (KHSG) wurde geregelt, dass Ergebnisse

zur Struktur-, Prozess- und Ergebnisqualität der stationären Gesundheitsversorgung von den Landesplanungsbehörden für Zwecke der qualitätsorientierten Krankenhausplanung Verwendung finden sollen. Es findet demnach eine Erweiterung der Verwendung der gesetzlich verpflichtenden Qualitätsmessungen statt, die neben *improvement* (z. B. Peer-Reviews, eigenständige qualitätsorientierte Umgestaltung der Versorgung durch die Leistungserbringer) nun auch im Rahmen von *accountability* (z. B. öffentliche Berichterstattung, Planungsentscheidungen) eingesetzt werden (Solberg et al. 1997, Berwick et al. 2003). Sowohl der Einsatz für *improvement* als auch für *accountability* sind Strategien, die das gleiche Ziel einer besseren Gesundheitsversorgung verfolgen, jedoch über andere Wirkmechanismen (Berwick et al. 2003). So gehen etwa Berwick et al. davon aus, dass durch die öffentliche Verfügbarkeit aussagekräftiger, leistungserbringerbezogener Qualitätsinformationen sowohl Patientinnen und Patienten als auch andere Entscheidungsträger (z. B. Landesplanungsbehörden) eine Selektion von Leistungserbringern mit höherer Qualität vornehmen. Dies wiederum sollte Leistungserbringern mit niedrigerer Qualität einen Anreiz geben, durch eigene Anstrengungen die Qualität ihrer Versorgung zu steigern. Auf der anderen Seite gehen Berwick et al. davon aus, dass durch die Rückmeldung von Qualitätsergebnissen an die Leistungserbringer selbst, diese weitergehende Einblicke in ihre Versorgungsqualität erhalten und damit in die Position versetzt werden, die Versorgung so umzugestalten, dass sie eine höhere Qualität erreicht (*improvement*). Das bisherige Vorgehen des Strukturierten Dialog kann damit eher letzterer Strategie zugeordnet werden, während die neueren Verwendungszwecke der Qualitätsmessungen der gesetzlich verpflichtenden Qualitätssicherung eher die erstgenannte Strategie verfolgen.

Berücksichtigt man diese Erweiterung der Verwendung von Qualitätsindikatoren für öffentliche Berichterstattung oder um Planungsentscheidungen (*accountability*), steigen auch die methodischen Anforderungen an Qualitätsindikatoren (Solberg et al. 1997, Chassin et al. 2010, Berwick et al. 2003, Gardner et al. 2018, Freeman 2002). Um im geschützten Raum in einen Dialog über Verbesserungsmöglichkeiten der Versorgungsqualität zu führen, genügen „Aufgreifkriterien". Um jedoch die Ergebnisse von indikatorbasierten Qualitätsmessungen für beispielsweise Planungs- und Vergütungsentscheidungen oder Patienteninformationen zu verwenden, müssen die Indikatorergebnisse aussagekräftiger sein. Sollen nämlich beispielsweise quantitative Indikatorergebnisse miteinander verglichen werden, müssen numerische Unterschiede zwischen Leistungserbringerergebnissen auch mindestens tendenziell Unterschiede[9] in der Qualität der Versorgung widerspiegeln.[10] Es genügt damit nicht mehr, diese als bloße Aufgreifkriterien der Versorgungsqualität zu verstehen. Stattdessen müssen sie auf Basis hoher Datenqualität zuverlässige Rückschlüsse auf die Versorgungsqualität zulassen. Entsprechend dieser gestiegenen Anforderungen an die datengestützte Qualitätssicherung ist ein Wandel im Verständnis von Qualitätsindikatoren in der gesetzlich verpflichtenden Qualitätssicherung erforderlich, weg von Aufgreifkriterien hin zu quantitativen Größen für Qualität (vgl. Solberg et al. 1997, IQTIG 2019a,

[9] Hier muss auch die Frage nach der Rolle statistischer Unsicherheit beantwortet werden.

[10] Dies kann auch als Frage des Skalenniveaus der Messung verstanden werden: Während Aufgreifkriterien eher Nominalskalenniveau (auffällig vs. unauffällig) zugeschrieben wird, setzt der Vergleich zweier Messwerte auf einem Indikator metrisches Skalenniveau voraus.

Carter 1989). Dies bedeutet nicht, dass Qualitätsmessungen grundsätzlich nicht als Aufgreifkriterien verwendet werden sollten, sondern dass deren Verwendung im Rahmen von *accountability* höhere methodische Anforderungen an die Messung stellt als im Rahmen von *improvement*.

Der bisherige Strukturierte Dialog mit Krankenhäusern im Rahmen der QSKH-RL weist auch Aspekte eines Peer-Review Verfahrens auf, in dem „auffällige Ergebnisse" zunächst in einem zweiten Prüfschritt im Dialog mit dem Leistungserbringer „geklärt werden". Diese sprachliche Trennung von quantitativem Messergebnis („rechnerisches Ergebnis") und Qualitätsbewertung im Rahmen des dialogischen Klärungsprozesses in der QSKH-RL könnte als implizites Verständnis von Qualitätsindikatoren als Aufgreifkriterien verstanden werden, deren numerischer Wert eher geringere Aussagekraft hat und eher die Funktion erfüllt, einen darauffolgenden Prüfschritt auszulösen.

Andererseits werden schon heute die quantitativen Indikatorergebnisse in den Strukturierten Qualitätsberichten der Krankenhäuser veröffentlicht. Dies bedeutet, dass den Indikatorergebnissen selbst auch eine Funktion (und Information) über die als Aufgreifkriterien für den Strukturierten Dialog hinaus zugeschrieben wird. Im Rahmen der Beauftragung zur Veröffentlichung von leistungserbringervergleichenden Übersichten über die Qualität in der stationären Versorgung (sog. G-BA-Qualitätsportal), soll das IQTIG unter anderem auch die Frage beantworten, ob metrische Ergebnisse mehrerer Indikatoren verdichtet werden sollen. Eine solche Verdichtung oder Aggregation von Indikatorergebnissen setzt voraus, dass die Indikatorergebnisse mehr als bloße Aufgreifkriterien darstellen und ihre numerischen Ergebnisse selbst eine Aussagekraft haben.

Des Weiteren ergibt sich ein Zielkonflikt zwischen dem bisherigen, ressourcenintensiven dialogischen Vorgehen im zweiten Schritt und dem Wunsch nach einer „Verschlankung des Verfahrens" (vgl. Beauftragung). Zukünftig wird die Anzahl an Qualitätsindikatoren eher steigen, da weitere Leistungsbereiche einschließlich Patientenbefragungen für die QS erschlossen werden. Wie in Abschnitt 2.1 erläutert, geht mit der Verlagerung von Spezifität von der indikatorbasierten Klassifikation hin zum nachgelagerten Prüfschritt auch eine Aufwandsverlagerung einher. So bedeutet eine größere Anzahl an Indikatoren, die nach dem hier beschriebenen Verständnis eher als Aufgreifkriterien mit eher niedriger Spezifität konstruiert sind, auch einen erhöhten Aufwand im nachgelagerten Prüfschritt.

In der Zusammenschau der gestiegenen (methodischen) Anforderungen an die gesetzliche Qualitätssicherung und dem Wunsch nach einem ressourcenschonenderen Vorgehen ergibt sich die Notwendigkeit, den Prozess der Qualitätsmessung stärker, auch mittels automatisierbarer statistischer Methoden, zu standardisieren und die aufwendigen dialogischen Anteile des Verfahrens weg von der Qualitätsbewertung hin zur Qualitätsförderung zu verlagern. Dies bedeutet jedoch nicht, dass Maßnahmen zur Steigerung der Versorgungsqualität (*improvement*) weniger wichtig werden, nur weil die Ergebnisse der Qualitätsmessungen zusätzlich im Kontext von *accountability* Verwendung finden. Wie später (siehe Kapitel 1) erläutert wird, sollten diese Maßnahmen jedoch von den Prozessen der Qualitätsmessung und -bewertung aus mehreren Gründen getrennt werden und ihren sinnvollen Einsatz bei der Qualitätsförderung (*improvement*) finden.

2.3 Strukturelle Voraussetzungen des Strukturierten Dialogs

Im Folgenden wird zunächst der Ist-Zustand der strukturellen Voraussetzungen für die Durchführung des Strukturierten Dialogs beschrieben. Der bisherige Strukturierte Dialog findet zwischen dem jeweiligen Leistungserbringer (Fachabteilung eines Krankenhausstandorts und dem abteilungsübergreifenden internen Qualitätsmanagement der Einrichtung) und der jeweiligen beauftragten Stelle (verantwortliche Einrichtungen der Landesebene bzw. dem IQTIG auf Bundesebene) statt.

Die beauftragten Stellen auf Bundes- und Landesebene richten zusätzlich Fachgruppen[11] mit fachkundigen ehrenamtlichen Expertinnen und Experten ein, um die Bewertung der Qualitätsindikatorergebnisse vorzunehmen und um Qualitätssicherungsmaßnahmen anzustoßen. Laut QSKH-RL werden die fachkundigen Mitglieder unter anderem von den Trägerorganisationen des G-BA entsendet. In Bezug auf die Qualifikation der Mitglieder werden ausschließlich Kenntnisse aus dem Bereich Qualitätsmanagement und der Qualitätssicherung gefordert. Auf welchem Gebiet die Mitglieder darüber hinaus fachkundig sein sollten, ist nicht näher beschrieben. Folglich ist unklar, über welche genaue Qualifikation die einzelnen Mitglieder verfügen sollen. Bundesweit existieren derzeit insgesamt 114 auf Landesebene und durchschnittlich 7 Fachgruppen pro beauftragte Stelle, die für die Umsetzung der Maßnahmen der externen stationären Qualitätssicherung zuständig sind (siehe Anhang, Kapitel 3, Tabelle 1). Die Anzahl der Mitglieder pro Fachgruppe variiert stark sowohl zwischen den Bundesländern als auch zwischen den Fachgruppen. Laut einer Befragung des G-BA aus dem Jahr 2012, an der alle Landesgeschäftsstellen für Qualitätssicherung sowie die Institution nach § 137a SGB V teilnahmen, schwankt die Anzahl der Fachgruppenmitglieder je nach Bundesland zwischen 3 und 18 Mitgliedern. Innerhalb einer Fachgruppe sind durchschnittlich 7 Mitglieder vertreten [(siehe Anhang, Kapitel 3, Tabelle 1; Winkler-Komp et al. (2014)].

Bei allen indirekten Verfahren existieren zusätzlich zu den Fachgruppen auf Landesebene auch Fachgruppen auf Bundesebene, die jedoch andere Aufgaben wahrnehmen und daher hier nicht thematisiert werden.

Auch das IQTIG richtet verschiedene Fachgruppen für die Umsetzung des Strukturierten Dialogs für die direkten Verfahren ein.

2.4 Durchführung des bisherigen Strukturierten Dialogs

Die Vorgaben in der QSKH-RL legen lediglich Eckpunkte des Ablaufs des Strukturierten Dialogs fest, sodass die konkrete Ausgestaltung der oben genannten drei Funktionen und hinsichtlich des Zeitraums, in welchem die Leistungserbringer eine Stellungnahme abzugeben haben, ein großer Gestaltungsspielraum existiert.

[11] Im Folgenden wird die Bezeichnung Fachgruppe aus der QSKH-RL verwendet, da sich die Analyse der Ausgangssituation auf den Strukturierten Dialog im Rahmen der QSKH-RL bezieht. Neben dem Begriff Fachgruppe existieren z. B. auf Ebene der Länder oder in anderen Richtlinien wie der DeQS-RL auch andere Bezeichnungen wie z. B. Arbeitsgruppe, Fachbereich, Fachkommission, Fachausschüsse usw. Ab Kapitel 4 wird für den Begriff Fachgruppe der Begriff Fachkommission gemäß DeQS-RL verwendet.

Abbildung 1 veranschaulicht schematisch den Ablauf des Strukturierten Dialogs. Weicht das auf Grundlage der übermittelten Daten für einen Krankenhausstandort errechnete Ergebnis vom bundesweit einheitlich festgelegten Referenzbereich ab (*rechnerische Auffälligkeit*), ist der Leistungserbringer darauf hinzuweisen oder es ist eine Stellungnahme vom Leistungserbringer einzuholen (§ 11 QSKH-RL). Auf einen Hinweis an den Leistungserbringer oder die Einholung einer Stellungnahme kann laut § 10 QSKH-RL verzichtet werden, wenn für den Standort nur ein Fall im Qualitätsindikator auftrat. Werden bei einem rechnerisch auffälligen Ergebnis keine Maßnahmen eingeleitet, ist der Grund dafür darzustellen und anzugeben (z. B. bei der Verwendung von Qualitätsindikatorensets oder von landesindividuellen Referenz- und Vertrauensbereichen, § 10 QSKH-RL).

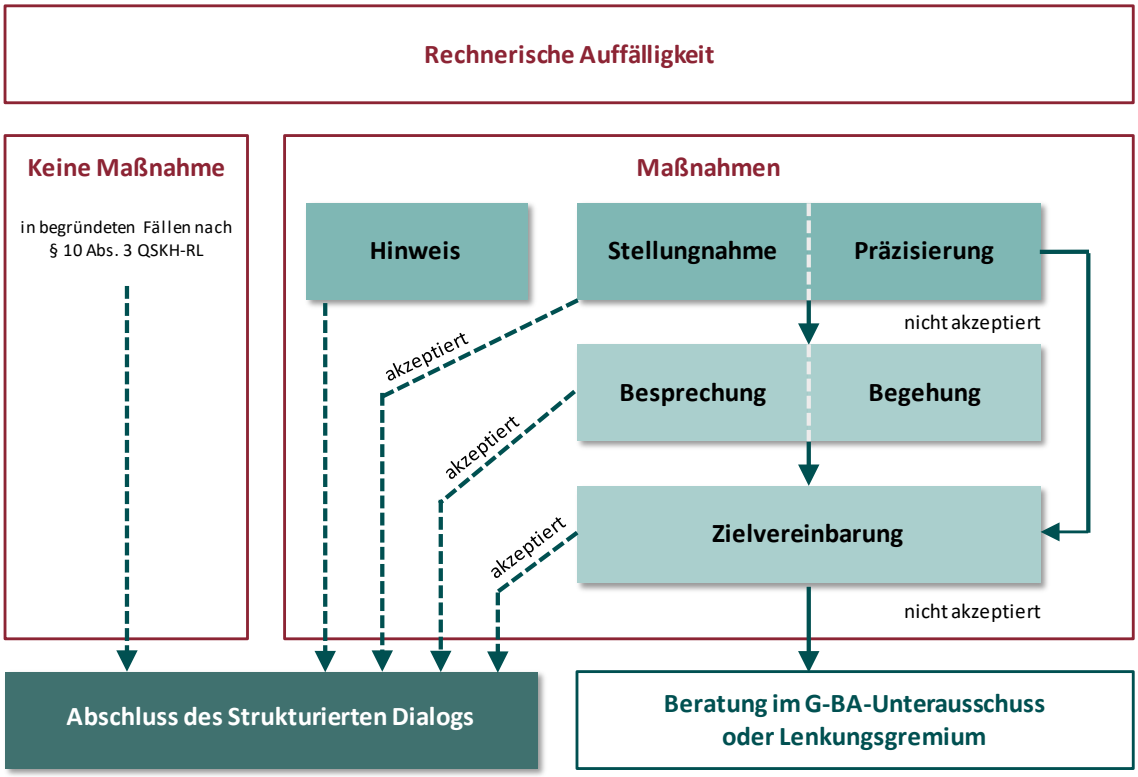

Abbildung 1: Schematische Darstellung von Maßnahmen im Rahmen des Strukturierten Dialogs

Welche Form oder welchen Inhalt eine Stellungnahme haben soll, ist nicht definiert. In § 12 QSKH-RL ist weiterhin geregelt, dass die Stellungnahme dahingehend geprüft werden soll, ob die dokumentierten Leistungen in der fachlich gebotenen Qualität erbracht wurden. Wie genau und nach welchen Kriterien diese Prüfung stattfinden soll, bleibt offen. In der Praxis werden mit unbekannter Häufigkeit Präzisierungsanfragen zu der ursprünglichen Stellungnahme des Leistungserbringers gestellt.

Am Ende des Strukturierten Dialogs werden alle auffälligen Ergebnisse in das bundeseinheitliche Einstufungsschema (siehe Anhang, Kapitel 3, Tabelle 2) eingeordnet. Es beinhaltet 7 Kategorien (AQUA 2013: 55). Eine Ausdifferenzierung mit diversen Ziffern führt in der Summe zu insgesamt

21 unterschiedlichen Kategorien.[12] Die Kategorien „N“ und „R“ des Einstufungsschemas beziehen sich auf rechnerisch unauffällige Ergebnisse, bei welchen kein Strukturierter Dialog stattfindet. Daher wird auf diese Kategorien im Folgenden nicht weiter eingegangen.

2.4.1 Einstieg in den Strukturierten Dialog

Der Strukturierte Dialog kann durch zwei Arten von rechnerisch auffälligen Ergebnissen ausgelöst werden.[13] Zum einen kann ein Leistungserbringer ein rechnerisch auffälliges Ergebnis in einem *Qualitätsindikator* aufweisen. Zum anderen führt auch ein rechnerisch auffälliges Ergebnis in einem *Auffälligkeitskriterium* zum Einstieg in den Strukturierten Dialog. Auffälligkeitskriterien sind Kennzahlen zur Messung der Dokumentationsqualität. Sie prüfen die Plausibilität, Vollständigkeit und Vollzähligkeit der Dokumentation zum Zwecke der Qualitätssicherung. Zur Bewertung dieser Auffälligkeitskriterien existieren im Kategorienschema (vgl. Anhang, Kapitel 3, Tabelle 2) zwei Ziffern (*U30* und *A40*, grün hinterlegt), die exklusiv im Kontext der Datenvalidierung zu verwenden sind.

In Abhängigkeit davon, ob sich das rechnerisch auffällige Ergebnis auf einen Qualitätsindikator oder auf ein Auffälligkeitskriterium bezieht, hat auch die Bewertung des Ergebnisses einen anderen Fokus. Handelt es sich um ein rechnerisch auffälliges Ergebnis in einem Qualitätsindikator, liegt der Fokus auf der Bewertung der Versorgungsqualität. Handelt es sich um ein rechnerisch auffälliges Ergebnis in einem Auffälligkeitskriterium, liegt der Fokus auf der Bewertung der Dokumentationsqualität. Wie bereits in Abschnitt 1.3 beschrieben liegt der Fokus dieses Berichts auf der Bewertung der Versorgungsqualität, daher wird im Folgenden auf die Bewertung der Dokumentationsqualität nicht weiter eingegangen.

Wird ein rechnerisch auffälliges Ergebnis in einen Qualitätsindikator festgestellt, hat die mit der Durchführung des Strukturierten Dialogs beauftragte Stelle (LQS oder IQTIG) gemäß QSKH-RL zwei Handlungsoptionen: Das Krankenhaus ist „unter Beschreibung des Sachverhalts und Bezeichnung des betroffenen Standorts auf die Auffälligkeit hinzuweisen oder [es ist] eine Stellungnahme innerhalb einer festzusetzenden Frist anzufordern“ (§ 11 QSKH-RL). Entscheidet sich die mit der Durchführung des Strukturierten Dialogs beauftragte Stelle (LQS oder IQTIG), das Krankenhaus zu einer schriftlichen Stellungnahme aufzufordern, hat der Leistungserbringer die Gelegenheit, in der schriftlichen Stellungnahme darzulegen, warum die Leistungen trotz rechnerisch auffälligem Ergebnis in der fachlich gebotenen Qualität erbracht worden sind (§ 12 Abs. 1 QSKH-RL).

2.4.2 Bewertung der Versorgungsqualität

Ob bei einem rechnerisch auffälligen Qualitätsindikatorergebnis zunächst nur ein Hinweis versendet oder eine schriftliche Stellungnahme angefordert wird, ist entscheidend für die Quali-

[12] Die resultierende Bezifferung (z. B. *A41: Hinweise auf Struktur- und Prozessmängel*) impliziert auch die verpflichtende Darstellung für die Veröffentlichung der Ergebnisse in den jährlichen Qualitätsberichten der Krankenhäuser.

[13] Unabhängig von einem rechnerisch auffälligen Ergebnis kann der Strukturierte Dialog im Rahmen der Zweiterfassung der Datenvalidierung eingeleitet werden.

tätsbeurteilung. Eine Qualitätsbeurteilung der Leistungserbringung wird nur nach Einholung einer schriftlichen Stellungnahme vorgenommen. Im Fall der Versendung eines Hinweises erfolgt keine Beurteilung des rechnerisch auffälligen Indikatorergebnisses. Bisher existieren keine Regularien dafür, anhand welcher Kriterien diese erste wichtige Entscheidung zu treffen ist.

Eine im Jahr 2012 stattgefundene Befragung des G-BA der Landesgeschäftsstellen für Qualitätssicherung gibt Auskunft darüber, auf welcher Grundlage die Landesebene entscheiden. Die Befragungsergebnisse zeigen, dass die Entscheidung der Fachgruppen, welche Maßnahme ergriffen wird, oft anhand sehr unterschiedlicher Kriterien und häufig anhand mehrerer Kriterien getroffen wird. In einigen Ländern erfolgt die Entscheidung anhand:

- der Relevanz des Indikators
- der statistischen Signifikanz von Abweichungen
- des Vorliegens einer Risikoadjustierung

Aus den Ergebnissen der Befragung des G-BA geht weder hervor, wie die Relevanz des Indikators bestimmt oder welches Signifikanzniveau gewählt wird noch inwieweit dies einheitlich zwischen den beauftragten Stellen erfolgt (Winkler-Komp et al. 2014).

Das IQTIG holt in den direkten QS-Verfahren für jedes rechnerisch auffällige Ergebnis (Qualitätsindikator und Auffälligkeitskriterium) eine schriftliche Stellungnahme ein.

Hinweise

Einen Überblick über den Anteil der eingesetzten Hinweise in den einzelnen Bundesländern gibt die Tabelle 3 im Anhang, Kapitel 3. Es ist ersichtlich, dass der Anteil an versandten Hinweisen an allen rechnerisch auffälligen Ergebnissen stark zwischen den Bundesländern variiert. In Bayern und Baden-Württemberg liegt der Anteil bei weit über 50 %, während in Bremen und im Saarland kaum Hinweise versandt werden. Bundesweit wird im Durchschnitt für mehr als ein Drittel aller rechnerisch auffälligen Indikatorergebnisse ein Hinweis (Kategorie „H“) an den Leistungserbringer versendet. Durch die fehlende Qualitätsbeurteilung wird das rechnerisch auffällige Ergebnis zwar nicht als unauffällig deklariert, dennoch kann die fehlende Qualitätsbeurteilung den Eindruck erwecken, dass das rechnerisch auffällige Ergebnis unauffällig sei. Tatsächlich ist es jedoch so, dass zu rechnerisch auffälligen Ergebnissen, bei welchen ein Hinweis versandt wurde, keinerlei Aussagen zu möglichen qualitativen Versorgungsproblemen gemacht werden können. Das bedeutet, dass in vielen Bundesländern für eine große Anzahl an rechnerisch auffälligen Ergebnissen keine Qualitätsbewertung vorgenommen wird. In den direkten Verfahren wird aufgrund der kleineren Anzahl an teilnehmenden Standorten und Abteilungen jedes rechnerisch auffällige Ergebnis mindestens durch die Anforderung einer Stellungnahme überprüft.

Schriftliche Stellungnahmen

Fordert die beauftragte Stelle eine schriftliche Stellungnahme vom Leistungserbringer an, wird diese dahingehend geprüft, ob die medizinischen Leistungen mit der fachlich gebotenen Qualität erbracht und valide dokumentiert wurden (§ 11 Abs. 1 QSKH-RL). Das bedeutet, dass anhand der schriftlichen Stellungnahmen auch eine Einstufung der Dokumentationsqualität stattfindet (vgl. Kategorie „D“ in Tabelle 2 im Anhang, Kapitel 3), die unabhängig von der bereits erwähnten

Bewertung der Auffälligkeitskriterien durchgeführt wird. Für die Bewertung der Versorgungsqualität ist die Einstufung der Dokumentationsqualität insofern relevant, da eine *echte* Qualitätsbewertung des Indikatorergebnisses im Hinblick auf die Qualität medizinischer Leistungen (Kategorie „A“ oder „U“) nur stattfinden kann, wenn die Angaben der Leistungserbringer korrekt sind. Wurden in der schriftlichen Stellungnahme des Leistungserbringers keine Angaben zu Dokumentationsfehlern gemacht, kann geprüft werden, ob die medizinische Leistung in der fachlich gebotenen Qualität erbracht wurde. Kann der Leistungserbringer in seiner Stellungnahme darlegen, dass sich das rechnerisch auffällige Ergebnis nicht durch ein Qualitätsdefizit begründet, kann das rechnerisch auffällige Ergebnis von der beauftragten Stelle als qualitativ unauffällig eingestuft werden (Kategorie „U“). Verbindliche Kriterien für diesen Entscheidungsprozess existieren bisher nicht.

Aus dem Bericht zum Strukturierten Dialog zum Erfassungsjahr 2015 geht hervor, dass der Anteil an schriftlichen Stellungnahmen bezogen auf alle rechnerisch auffälligen Ergebnisse bei 61,8 % lag. Das entspricht 9.797 schriftlichen Stellungnahmen deutschlandweit, die von den LQS und vom IQTIG angefordert, von den Leistungserbringern erstellt und anschließend von den beauftragten Stellen und ihren Fachgruppen bewertet wurden (IQTIG 2017a: 54). Obwohl die schriftliche Stellungnahme die Grundlage für die Bewertung der Versorgungsqualität bildet, existieren hier ebenso wenig bundeseinheitliche Vorgaben, die die Leistungserbringer bei der Erstellung der schriftlichen Stellungnahmen einzuhalten haben. Wie detailliert die schriftliche Stellungnahme ausfällt, liegt in der Verantwortung des jeweiligen Leistungserbringers und beeinflusst den Aufwand und die Bewertung der Indikatorergebnisse durch die Fachgruppe. Gehen beispielsweise Stellungnahmen mit ungenauen, unzureichenden oder gar fehlenden Angaben ein, kann die beauftragte Stelle eine Präzisierung der Stellungnahme einholen, um weitergehende Informationen zu erhalten, die dann ggf. eine abschließende Bewertung erst ermöglichen. Allerdings ist der Umgang mit Präzisierungen in den Rahmenvorgaben der Richtlinie nicht verankert. Die Einholung einer Präzisierung sowie die damit zusammenhängende erneute Prüfung und Bewertung des gesamten Sachverhalts beinhaltet einen hohen Arbeitsaufwand für die beauftragte Stelle. Wie häufig Präzisierungen eingeholt werden ist außerdem nicht bekannt und wird bisher nicht berichtet. Folglich existieren auch für Präzisierungen weder auf Bundes- noch auf Landesebene inhaltliche oder formale einheitliche Kriterien.

Besprechungen[14] und Begehungen im Rahmen der Qualitätsbewertung

Laut QSKH-RL können auch weiterführende Maßnahmen veranlasst werden, wenn trotz Stellungnahme noch Zweifel bestehen. Bleiben also anhand der schriftlichen Stellungnahme und einer möglichen Präzisierung noch inhaltliche Fragen offen, sodass eine abschließende Bewertung des rechnerisch auffälligen Ergebnisses nicht vorgenommen werden kann, können weiterführende Maßnahmen eingeleitet werden. Weiterführende Maßnahmen, die als Grundlage für die Qualitätsbeurteilung dienen können, sind z. B. Besprechungen mit dem Leistungserbringer oder auch Vor-Ort-Begehungen. Letztere sind allerdings lediglich mit *Einverständnis des Krankenhauses* möglich (§ 12 Abs. 3 QSKH-RL). Ein bundesweit einheitliches Vorgehen zum Einsatz

[14] Die Begriffe Besprechungen und kollegiales Gespräch werden synonym verwendet.

und zur Durchführung dieser weiterführenden Maßnahmen mit anschließender Bewertung existiert nicht. Deutschlandweit fanden in allen QS-Verfahren zum Erfassungsjahr 2015 insgesamt 282 Besprechungen und 19 Begehungen statt (IQTIG 2017a: 54).

Aus den Erfahrungen des IQTIG ist zu berichten, dass diese Maßnahmen für die beauftragte Stelle, die entsprechende Fachgruppe und auch für den Leistungserbringer sehr ressourcenintensiv sind. In der Praxis wird nicht nur bei verbleibenden Zweifeln nach Sichtung der Stellungnahmen, sondern auch bei wiederholt rechnerisch auffälligen Ergebnissen eine Begehung durchgeführt. Der direkte Kontakt zu den Krankenhäusern in einem kollegialen Gespräch oder während einer Begehung ist eine effektive, allerdings auch eine ressourcenintensive Maßnahme im Strukturierten Dialog.

Prozess der Entscheidungsfindung

Die Grundlage für eine abschließende Bewertung des Indikatorergebnisses bildet bei der Beurteilung von rechnerischen Auffälligkeiten in der Regel die schriftliche Stellungnahme des Leistungserbringers. Für diese finale Entscheidung, ob ein rechnerisch auffälliges Indikatorergebnis als qualitativ auffällig oder unauffällig eingestuft wird, gibt es keine Vorgaben oder Kriterien in der QSKH-RL.

In der Befragung der Landesgeschäftsstellen für Qualitätssicherung zeigte sich, dass die Landesgeschäftsstellen ihre Fachgruppen in unterschiedlichem Umfang in die Bewertung des Indikatorergebnisses anhand der schriftlichen Stellungnahmen einbinden (Winkler-Komp et al. 2014). Einerseits gibt es Landesgeschäftsstellen, die ihren Fachgruppen nicht immer alle schriftlichen Stellungnahmen der Leistungserbringer zur Bewertung vorlegen. Beispielsweise bindet eine Landesgeschäftsstelle die Fachgruppen beim erstmaligen Vorliegen von Dokumentationsfehlern nicht in die Bewertung ein. Andererseits geht aus der Befragung hervor, dass die meisten Landesgeschäftsstellen ihren Fachgruppen keine Bewertungsvorschläge für das Indikatorergebnis zusammen mit den schriftlichen Stellungnahmen vorlegen, sodass die Bewertung direkt in der Sitzung vorgenommen wird. Es wird zudem beschrieben, wie der Prozess der Entscheidungsfindung im Rahmen der anschließenden Bewertung von Stellungnahmen (qualitativ auffälliges oder unauffälliges Ergebnis) auf der Landesebene getroffen wird. Am häufigsten wird darüber von den beauftragten Stellen im Konsens entschieden (siehe Tabelle 4 im Anhang, Kapitel 3). Ob die Landesgeschäftsstellen Informationen, die nicht Bestandteil der schriftlichen Stellungnahmen sind, für die Bewertung der Indikatorergebnisse heranziehen, ist nicht bekannt.

2.4.3 Bewertung der Dokumentationsqualität von Qualitätsindikatoren

Wird aufgrund eines rechnerisch auffälligen Ergebnisses eine schriftliche Stellungnahme angefordert, wird diese neben der inhaltlichen Überprüfung der Angaben zur medizinischen Versorgung auch dahingehend geprüft, ob der Leistungserbringer Angaben gemacht hat, die er aufgrund einer fehlerhaft vorgenommenen Dokumentation nicht zu verantworten hat. Hierbei ist zwischen zwei Arten der „Entlastung“ des Leistungserbringers aufgrund der Angabe von Dokumentationsfehlern zu unterscheiden. Leistungserbringer können im Rahmen der schriftlichen Stellungnahmen einerseits das Vorhandensein von Dokumentationsfehlern anführen, die im

Hinblick auf eine mögliche qualitative Auffälligkeit zu einer indirekten Entlastung führen, da in der Folge keine Qualitätsbeurteilung vorgenommen werden kann. Das Bewertungsschema sieht für diese Konstellationen vor, das rechnerisch auffällige und potenziell qualitativ auffällige Ergebnis in die Kategorie „D“ („Bewertung nicht möglich wegen fehlerhafter Dokumentation“) einzustufen. Bundesweit wurden für das Erfassungsjahr 2015 durchschnittlich 10,4 % der rechnerisch auffälligen Indikatorergebnisse aufgrund nachträglicher Angaben von Dokumentationsfehlern nicht bewertet (IQTIG 2017a: 64 ff.).

Andererseits können derzeit rechnerisch auffällige Ergebnisse als qualitativ unauffällig bewertet werden und damit direkt zu einer Entlastung des Leistungserbringers führen, wenn die beauftragte Stelle nach der Prüfung der Stellungnahme resümiert, dass vereinzelt Dokumentationsprobleme vorlagen. Bundesweit wurden für das Erfassungsjahr 2015 durchschnittlich 7,4 % der Ergebnisse aufgrund vereinzelter Dokumentationsprobleme als unauffällig eingestuft (Kategorie „U33“). Die Spannweite der Häufigkeit von Kategorie „U33“ reicht in Abhängigkeit vom Bundesland von 0 % bis 24,1 % (IQTIG 2017b: 52 ff.).

Fasst man die Kategorien „D“ und „U33“ zusammen, zeigt sich, dass in Deutschland im Jahr 2015 durchschnittlich 17,9 % aller rechnerisch auffälligen Ergebnisse aufgrund von nachträglich angegebenen Dokumentationsproblemen entweder indirekt oder direkt zu einer Entlastung des Leistungserbringers führten (IQTIG 2017b: 52 ff.).

2.4.4 Qualitätsförderung

In der Beauftragung zur Weiterentwicklung des Strukturierten Dialogs wird die Qualitätsförderung von Krankenhausleistungen als eindeutiges Ziel des Strukturierten Dialogs verstanden. In der QSKH-RL werden jedoch keine konkreten Maßnahmen zur Qualitätsförderung unter den *Maßnahmen der Externen stationären Qualitätssicherung* (QSKH-RL, Abschnitt B) genannt. Aufgrund der notwendigen und engen Abstimmung zwischen beauftragter Stelle und dem Leistungserbringer können vor allem Besprechungen, Begehungen und Zielvereinbarungen als qualitätsfördernde Maßnahmen und weniger als Instrumente der Qualitätsbewertung verstanden werden.

Besprechungen und Begehungen im Rahmen der Qualitätsförderung

Im Hinblick auf *Besprechungen* legt die QSKH-RL u. a. Folgendes fest: „Eine Besprechung dient der Aufklärung von Zweifeln und der erforderlichen, ggf. vom Krankenhaus erbetenen, Beratung“ (§ 12 Abs. 2 QSKH-RL). Demnach können gemeinsam qualitätsverbessernde Maßnahmen erarbeitet werden, insbesondere, wenn vonseiten des Leistungserbringers eine Beratung erwünscht und eine Zusammenarbeit mit der beauftragten Stelle gewollt ist. Nach einer Begehung, deren Fokus auf der Prüfung von Qualitätsmängeln liegt, soll eine Besprechung angeschlossen werden. Es kann davon ausgegangen werden, dass Besprechungen und Begehungen im Vergleich zur Einholung einer schriftlichen Stellungnahme beim Leistungserbringer durch den direkten Austausch eine stärkere qualitätsfördernde Funktion erfüllen. Da eine abschließende Bewertung des Indikatorergebnisses zum Zeitpunkt der Begehung jedoch ggf. noch nicht stattgefunden haben muss, ist zu vermuten, dass die noch ausstehende Bewertung möglicherweise

einen ungünstigen Einfluss auf die qualitätsfördernden Maßnahmen hat. Wird in diesem Zusammenhang das Ergebnis eines Leistungserbringers in einem Qualitätsindikator als qualitativ auffällig eingestuft, kann der in der Besprechung oder Begehung gemeinsam erkannte Verbesserungsbedarf in einer Zielvereinbarung festgehalten werden (§ 12 Abs. 2 QSKH-RL).

Besprechungen und Begehungen stellen für alle Beteiligten einen hohen Aufwand dar. Es ist anzunehmen, dass sie aus diesem Grund im Vergleich zu schriftlichen Stellungnahmen relativ selten angewandt werden. Bundesweit wurden im Jahr 2016 zum Erfassungsjahr 2015 282 Besprechungen und 19 Begehungen durchgeführt (IQTIG 2017a: 54).

Zielvereinbarungen

In einer Zielvereinbarung sollen qualitätsfördernde Maßnahmen vereinbart werden, die dem Leistungserbringer helfen sollen, sein Qualitätsproblem zu beheben. Die gemeinsame Erarbeitung von Zielen zur Qualitätsverbesserung und das schriftliche Festhalten einer diesbezüglichen Vereinbarung (§ 12 Abs. 2 QSKH-RL) ist ein wesentliches Element der Qualitätsförderung im Strukturierten Dialog. Zum Erfassungsjahr 2015 wurden 1.121 Ziele zwischen den beauftragten Stellen und Leistungserbringern vereinbart. Das entspricht einem Anteil an 7,1 % an allen rechnerisch auffälligen Ergebnissen. Da bundesweit deutlich mehr Zielvereinbarungen als Besprechungen und Begehungen (insgesamt 301) stattgefunden haben, kann geschlussfolgert werden, dass Zielvereinbarungen teilweise ausschließlich auf Basis schriftlicher Stellungnahmen abgeschlossen werden (IQTIG 2017a: 54).

2.5 Biometrische Analyse der Heterogenität in der Vorgehensweise und Bewertung im Strukturierten Dialog

Im Folgenden soll die Heterogenität in der Einstufung von Leistungserbringern durch die für die Durchführung des Strukturierten Dialogs beauftragten Stellen quantitativ analysiert werden. Speziell sollen Unterschiede zwischen den Bewertungsstellen bei der Einleitung eines Stellungnahmeverfahrens und bei der Qualitätsbewertung anhand der Daten des Strukturierten Dialogs zum Erfassungsjahr 2016 quantifiziert werden. Auch wenn die jährlichen Berichte zum Strukturierten Dialog hinreichend belegen, dass es etwa im Anteil an versendeten Hinweisen Unterschiede zwischen den Bewertungsstellen gibt (vgl. IQTIG (2017a), IQTIG (2018a)), bleibt zum einen die Frage, wie sich diese Unterschiede zusammenfassend quantifizieren lassen und zum anderen, wie stark sich etwaige Unterschiede zwischen den Bewertungsstellen auch durch Unterschiede in der Zusammensetzung der Leistungserbringer in den jeweiligen Bundesländern erklären lassen. Da in den Bewertungsstellen für jedes QS-Verfahren unterschiedliche Fachgruppen existieren, welche die Bewertung von Stellungnahmen vornehmen, ist auch innerhalb einer beauftragten Stelle Heterogenität bei der Bewertung von Stellungnahmen zwischen verschiedenen QS-Verfahren zu erwarten. Im Folgenden soll zunächst jedoch nur die Heterogenität zwischen den Bewertungsstellen exemplarisch anhand ausgewählter Qualitätsindikatoren analysiert werden.

Analog zur Risikoadjustierung bei Qualitätsindikatoren muss auch bei der Analyse der Unterschiede im Strukturierten Dialog der beauftragten Stellen gefragt werden: Inwiefern lassen sich

die beobachteten Unterschiede beispielsweise im Anteil qualitativ auffälliger Ergebnisse zwischen den Bewertungsstellen durch tatsächliche Qualitätsunterschiede der zugeordneten Leistungserbringer erklären und wie viel davon geht zurück auf eine heterogene Bewertungsweise der beauftragten Stellen?

Für die Analyse der Unterschiede zwischen den Bewertungsweisen der beauftragten Stellen sollen folgende Fragestellungen beantwortet werden:

- **Fragestellung 1**: Wie groß ist die bewertungsstellenabhängige Heterogenität bezüglich der Aufnahme eines Stellungnahmeverfahrens bei rechnerischer Auffälligkeit eines Leistungserbringers?
- **Fragestellung 2**: Wie groß ist die bewertungsstellenabhängige Heterogenität bezüglich der Bewertung als qualitativ auffällig nach Aufnahme eines Stellungnahmeverfahrens?

Diese Fragestellungen sind jeweils unter Berücksichtigung der rechnerischen Ergebnisse der Leistungserbringer in ihrem Bundesland zu beantworten, um auszuschließen, dass sich die beobachteten Unterschiede durch Unterschiede zwischen den Leistungserbringern erklären. Es geht also darum, zu charakterisieren, wie die jeweiligen Bewertungsstellen *vergleichbare rechnerische Ergebnisse* bewerten.[15]

Zur Beantwortung dieser Fragen wurde eine Methodik entwickelt, die mithilfe von sog. gemischten logistischen Regressionsmodellen (vgl. z. B. Agresti (2013), Stroup (2013)) in der Lage ist, den Einfluss von Bewertungsstelle und Leistungserbringerergebnis modellbasiert zu trennen. Im Anhang wird diese Methodik in Kapitel 4 ausführlich beschrieben. Dabei werden die Bewertungsergebnisse des Strukturierten Dialogs mit den rechnerischen Ergebnissen der Leistungserbringer verknüpft und ausgewertet. Die Methodik eignet sich allgemein zur Messung von Heterogenität in den Bewertungen des Strukturierten Dialogs und kann künftig auch im Rahmen weiterer Evaluationen verwendet und verallgemeinert werden.

Im konkreten Fall wurde die Methodik auf exemplarisch ausgewählte Qualitätsindikatoren und QS-Verfahren angewendet. Um ein aussagekräftiges Analyseergebnis zu erhalten, wurden für diese Analyse drei Qualitätsindikatoren ausgewählt, für welche zum Erfassungsjahr 2016 eine besonders hohe absolute Zahl an Stellungnahmen zu verzeichnen war. Dabei wurden jeweils ein Prozess-, Indikations- und Ergebnisindikator ausgewählt (vgl. Tabelle 2), die aus den QS-Verfahren bzw. Auswertungsmodulen *Herzschrittmacher-Implantation*, *Hüftendoprothesenversorgung* und *Pflege: Dekubitusprophylaxe* stammen.

[15] Dabei dienen die rechnerischen Ergebnisse der Leistungserbringer für die Analyse der qualitativ auffälligen Bewertungen (Fragestellung 2) lediglich als Approximation, um tatsächliche Qualitätsunterschiede zwischen den Leistungserbringern zu modellieren.

Tabelle 2: Indikatorenauswahl für exemplarische Analyse der Heterogenität des Strukturierten Dialogs. Alle Angaben beziehen sich auf den Strukturierten Dialog zum Erfassungsjahr 2016.

QI-ID	Modul-Kürzel	QI-Bezeichnung	Bewertungsart	Anz. Standorte	Anz. rechn. auff. Standorte	Anz. Stellungn.	Anz. qual. auff. Standorte
52010	DEK	Alle Patienten mit mindestens einem stationär erworbenen Dekubitalulcus Grad/Kategorie 4	Sentinel Event	1867	544	477	98
54003	HEP	Präoperative Verweildauer bei endoprothetischer Versorgung einer hüftgelenknahen Femurfraktur	Ratenbasiert	1231	625	384	134
54139	09/1	Leitlinienkonforme Indikation bei bradykarden Herzrhythmusstörungen	Ratenbasiert	1106	498	279	43

Eine detaillierte Darstellung der Ergebnisse für jeden der drei ausgewählten Qualitätsindikatoren ist im Anhang unter Abschnitt 4.2 zu finden. Beispielhaft werden die Ergebnisse der Analyse zum Indikator „Präoperative Verweildauer bei endoprothetischer Versorgung einer hüftgelenknahen Femurfraktur“ anhand von Abbildung 2 dargestellt. Dargestellt wird für Fragestellung 1 (oben) die modellbasierte Wahrscheinlichkeit P_{lb}, dass von einem rechnerisch auffälligen Leistungserbringer l durch Bewertungsstelle b eine Stellungnahme angefordert wird. Jedem Punkt in der Grafik entspricht eine Bewertungsstelle. Dargestellt wird für Fragestellung 2 (unten) die modellbasierte Wahrscheinlichkeit, dass ein Leistungserbringer l nach Anfordern einer Stellungnahme durch Bewertungsstelle b als qualitativ auffällig eingestuft wird. Dabei wird in den vier Spalten der Grafik dargestellt, wie sich die Wahrscheinlichkeiten für Fragestellung 1 und 2 in Abhängigkeit des rechnerischen Ergebnisses des Leistungserbringers l verhalten. Beispielsweise zeigen die Bewertungswahrscheinlichkeiten zu Fragestellung 1 von Leistungserbringern im 1. Quartil der Ergebnisse (links) eine größere bewertungsstellenabhängige Streuung als bei Leistungserbringern im 4. Quartil der Ergebnisse (rechts).

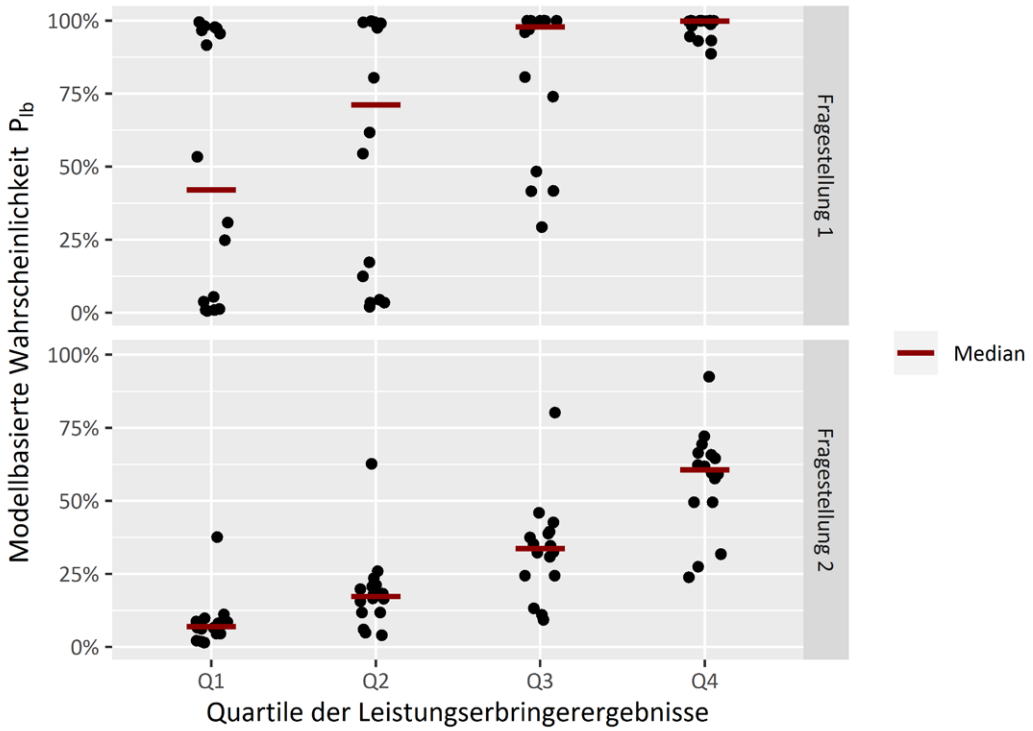

Abbildung 2: Beispielhafte Darstellung der Analyseergebnisse zum Indikator „Präoperative Verweildauer bei endoprothetischer Versorgung einer hüftgelenknahen Femurfraktur" (QI-ID 54003)

Die wichtigsten Erkenntnisse, die sich anhand der Analyseergebnisse belegen lassen, sind dabei Folgende:

Zunächst lässt sich zeigen, dass die Wahrscheinlichkeit für das Anfordern einer Stellungnahme abhängig sowohl von der Bewertungsstelle als auch vom Ausmaß der Überschreitung des indikatorspezifischen Referenzbereichs durch den Leistungserbringer ist. Extreme Überschreitungen des Referenzbereichs führen häufiger zum Anfordern einer Stellungnahme als knappe Überschreitungen des Referenzbereichs. Analoge Beobachtungen lassen sich nach dem Beginn des Stellungnahmeprozesses auch für die Wahrscheinlichkeit einer Bewertung eines Leistungserbringers als qualitativ auffällig machen. Auch hier unterscheiden sich die Bewertungswahrscheinlichkeiten je nach Bewertungsstelle und Ausmaß der Überschreitung des Referenzbereichs. Leistungserbringer, die den Referenzbereich deutlich überschreiten, werden häufiger als qualitativ auffällig bewertet als Leistungserbringer, die den Referenzbereich nur knapp verfehlen.

Zu den Eingangs formulierten Fragestellungen 1 und 2 nach der Heterogenität der Bewertungsergebnisse lässt sich zeigen, dass eine stark ausgeprägte bewertungsstellenabhängige Heterogenität unabhängig von den numerischen Indikatorergebnissen der Leistungserbringer in den jeweiligen Bundesländern besteht. Im Detail ist dabei anzumerken, dass diese Effekte je nach Qualitätsindikator und nach Fragestellung (Anfordern von Stellungnahmen (1) oder Bewertung als qualitativ auffällig (2)) unterschiedlich ausgeprägt sind, vgl. Anhang, Abschnitt 4.2. So ist bei den beiden ratenbasierten Indikatoren „Präoperative Verweildauer bei endoprothetischer Ver-

sorgung einer hüftgelenknahen Femurfraktur“ und „Leitlinienkonforme Indikation bei bradykarden Herzrhythmusstörungen“ die Heterogenität im Versenden von Hinweisen/Anfordern von Stellungnahmen (Fragestellung 1) deutlich ausgeprägter als beim Sentinel-Event-Indikator zu einem stationär erworbenen Dekubitalulcus Grad 4, welcher wiederum eine höhere Heterogenität bei der Bewertung als qualitativ auffällig (Fragestellung 2) zeigt als die beiden anderen Indikatoren.

Eine weitere Erkenntnis bezüglich der Heterogenität der Bewertungen ist, dass die größten Unterschiede zwischen den beauftragten Stellen bei der Bewertung von Leistungserbringern auftreten, deren numerische Indikatorergebnisse den Referenzbereich nur knapp überschreiten. Dies lässt sich anhand der Abbildung 2 (oben) bzw. anhand der Abbildungen 2 bis 4 im Anhang unter Abschnitt 4.2 nachvollziehen.

Für das Ziel der Vereinheitlichung des Strukturierten Dialogs lässt sich aus der stark ausgeprägten Heterogenität bezüglich Fragestellung 1 ableiten, dass eine einheitliche Handhabung beim Versenden von Hinweisen bzw. Anfordern von Stellungnahmen durch die beauftragten Stellen als ausschlaggebender Schritt anzusehen ist – im Hinblick auf die daraus resultierenden Unterschiede in den Ergebnissen. Insbesondere bei Leistungserbringern mit Indikatorergebnissen, die den Referenzbereich nur knapp verletzen, sind einheitliche Regelungen erforderlich, wann eine Stellungnahme verpflichtend anzufordern ist. So könnte beispielsweise durch die Berücksichtigung statistischer Unsicherheit bei der Auffälligkeitseinstufung eine Regelung getroffen werden, die diesen besonders heterogen gehandhabten Wertebereich numerischer Indikatorergebnisse einheitlich regelt. Es ist wichtig darauf hinzuweisen, dass an diesem Punkt die beiden Auftragsaspekte „Vereinheitlichung“ und „Effizienzsteigerung“ des Strukturierten Dialogs zusammen zu denken sind. Eine große Anzahl an rechnerisch auffälligen Leistungserbringern führt bei vielen Bewertungsstellen zu einer Priorisierung: Von Leistungserbringern, die den Referenzbereich stark überschreiten, wird häufiger eine Stellungnahme angefordert als von Leistungserbringern, die den Referenzbereich nur schwach überschreiten. Gelingt es, durch eine veränderte quantitative Auffälligkeitseinstufung eine sinnvolle Reduktion der anzufordernden Stellungnahmen zu erreichen, können mehr Ressourcen zur Bearbeitung der einzelnen Stellungnahmen zur Verfügung stehen.

Eine weitere Konsequenz der vorliegenden Analyseergebnisse ist, dass die Vereinheitlichung des Strukturierten Dialogs nicht auf die Frage der Handhabung von Hinweisen zu reduzieren ist: In den Ergebnissen zu Fragestellung 2 wurde ebenfalls eine deutliche Heterogenität bei der qualitativen Bewertung von Stellungnahmen festgestellt, die belegt, dass auch hier Vereinheitlichungsbedarf besteht.

2.6 Biometrische Analyse zu Aufwand und Effizienz des Strukturierten Dialogs

Auftragsgegenstand der Weiterentwicklung des Strukturierten Dialoges ist eine „Optimierung der Effizienz des Verfahrens“ (G-BA 2018a). Ausgangspunkt für eine Effizienz-Steigerung muss dabei eine Feststellung des Ist-Zustandes der Effizienz sein. Im folgenden Abschnitt wird exemplarisch anhand des QS-Verfahrens *Hüftendoprothesenversorgung* (HEP) dargestellt, welcher Ressourcenaufwand – gemessen anhand der Anzahl zu führender Strukturierter Dialoge – in der aktuellen Form der rechnerischen Auffälligkeitseinstufung für den Strukturierten Dialog entsteht. Diesem Aufwand werden die durch den Strukturierten Dialog identifizierten Qualitätsdefizite als „Nutzen“ des Strukturierten Dialoges gegenübergestellt. Dies entspricht einem eindimensionalen Verständnis von Effizienz als Verhältnis von Aufwand und Nutzen.

Durch diese deskriptive Analyse können Probleme und Einsparungspotenzial in der gegenwärtigen Methodik zur rechnerischen Auffälligkeitseinstufung gezielt beschrieben und adressiert werden. Die hier vorgestellten Zahlen sind in weiten Teilen bekannt aus dem jährlichen Bericht zum Strukturierten Dialog (IQTIG 2018a) – exemplarisch werden die Analysen für das Erfassungsjahr 2017 durchgeführt, welche das aktuellste Erfassungsjahr ist, für welches die Ergebnisse des Strukturierten Dialogs vorliegen. Ein besonderer Fokus der folgenden Aufwandsanalyse ist die Darstellung von rechnerischen und qualitativen Auffälligkeiten in Abhängigkeit von der Art des Referenzbereiches und der dem Indikatorergebnis zugrunde liegenden Fallzahl. Die Fallzahl ist dabei die Kardinalität der Grundgesamtheit des Indikators (n), welche zur Berechnung eines Standortergebnisses verwendet wird. Je nach Qualitätsindikator ist dies die Anzahl an Patientinnen und Patienten, die Anzahl durchgeführter Prozeduren oder die Anzahl stationärer Behandlungsfälle. Als Art des Referenzbereiches wird zwischen Qualitätsindikatoren mit festem Referenzbereich, Qualitätsindikatoren mit perzentilbasiertem Referenzbereich und Sentinel-Event-Indikatoren unterschieden.

2.6.1 Aufwandsüberblick QS-Verfahren *Hüftendoprothesenversorgung*

Im QS-Verfahren *Hüftendoprothesenversorgung* (HEP) gab es im Erfassungsjahr 2017 ca. 1280 Standorte, die Daten zu hüftendoprothetischen Eingriffen übermittelt haben. In 14 Qualitätsindikatoren gab es auf Standortebene ca. 17.000 zu bewertende Indikatorergebnisse. Mehr als 2.000 dieser Ergebnisse wurden als rechnerisch auffällig eingestuft. Das HEP-Verfahren verursacht, gezählt über alle QSKH-Verfahren, ca. 19 % aller rechnerischen Auffälligkeiten. Von diesen rechnerisch auffälligen Ergebnissen wurden 223 Ergebnisse im Strukturierten Dialog als qualitativ auffällig bewertet.

2.6.2 Rechnerische Auffälligkeiten je Qualitätsindikator

Tabelle 3 zeigt, wie sich die rechnerisch auffälligen Ergebnisse des HEP-Verfahrens auf die 14 Qualitätsindikatoren verteilen.

Tabelle 3: Rechnerisch auffällige Standorte im HEP-Verfahren für alle 14 Qualitätsindikatoren. Daten aus dem Erfassungsjahr 2017

QI-ID	QI-Titel	Art des Referenzbereiches	Anzahl Standorte	rechn. auff. Standorte	Anteil
54003	Präoperative Verweildauer bei endoprothetischer Versorgung einer hüftgelenknahen Femurfraktur	Fester Referenzwert	1226	527	43 %
54013	Todesfälle während des akut-stationären Aufenthaltes bei geringer Sterbewahrscheinlichkeit	Sentinel-Event	1251	320	25,6 %
54002	Indikation zum Hüftendoprothesen-Wechsel bzw. -Komponentenwechsel	Fester Referenzwert	1149	193	16,8 %
54017	Allgemeine Komplikationen bei Hüftendoprothesen-Wechsel bzw. -Komponentenwechsel	95. Perzentil	1149	146	12,7 %
54010	Beweglichkeit bei Entlassung	Fester Referenzwert	1193	140	11,7 %
54120	Spezifische Komplikationen bei Hüftendoprothesen-Wechsel bzw. -Komponentenwechsel	95. Perzentil	1149	133	11,6 %
54001	Indikation zur elektiven Hüftendoprothesen-Erstimplantation	Fester Referenzwert	1195	104	8,7 %
54016	Allgemeine Komplikationen bei elektiver Hüftendoprothesen-Erstimplantation	95. Perzentil	1195	102	8,5 %
54019	Spezifische Komplikationen bei elektiver Hüftendoprothesen-Erstimplantation	95. Perzentil	1195	92	7,7 %
54015	Allgemeine Komplikationen bei endoprothetischer Versorgung einer hüftgelenknahen Femurfraktur	95. Perzentil	1228	87	7,1 %
54012	Verhältnis der beobachteten zur erwarteten Rate (O/E) an Patienten mit Gehunfähigkeit bei Entlassung	95. Perzentil	1267	82	6,5 %
54018	Spezifische Komplikationen bei endoprothetischer Versorgung einer hüftgelenknahen Femurfraktur	95. Perzentil	1228	76	6,2 %
10271	Verhältnis der beobachteten zur erwarteten Rate (O/E) an Hüftendoprothesen-Wechsel bzw. -Komponentenwechsel im Verlauf	95. Perzentil	1279	69	5,4 %
54004	Sturzprophylaxe	Fester Referenzwert	1272	67	5,3 %
Alle	**-**		**16976**	**2138**	**12,6 %**

Unter den 5 Indikatoren mit den meisten rechnerischen Auffälligkeiten befinden sich drei Indikatoren mit festem Referenzbereich, darunter der Qualitätsindikator „Präoperative Verweildauer bei endoprothetischer Versorgung einer hüftgelenknahen Femurfraktur“, der im Erfassungsjahr 2017 als Qualitätsindikator mit besonderem Handlungsbedarf eingestuft wurde. Etwa 25 % aller rechnerischen Auffälligkeiten dieses QS-Verfahrens fallen allein auf diesen Qualitätsindikator. Auch der Sentinel-Event-QI „Todesfälle während des akut-stationären Aufenthaltes bei geringer Sterbewahrscheinlichkeit“ löst mehr als 300 Strukturierte Dialoge aus. Alle anderen Qualitätsindikatoren bewegen sich im Bereich von 67 bis 193 rechnerisch auffälligen Standortergebnissen.

Der Anteil der rechnerisch auffälligen Standorte bei Indikatoren mit perzentilbasierten Referenzwerten (in allen Fällen gilt in HEP das 95. Perzentil der Standortergebnisse als Referenzwert), liegt immer deutlich über dem nominell angestrebten Wert von 5 % auffälligen Standorten. Dies ist auf die besondere Art der Perzentilberechnung für den Referenzwert zurückzuführen, die den Perzentil-Referenzbereich nur auf Standorten mit wenigstens 20 Nenner-Fällen berechnet.[16] So ist beispielsweise die Anzahl der rechnerischen Auffälligkeiten bei den Komplikationsindikatoren zu Hüftendoprothesen-Wechseln (QI 54120 und QI 54017) mehr als doppelt so hoch wie das nominell suggerierte Niveau von 5 %. Diese Indikatoren haben einen sehr hohen Anteil an Standorten mit weniger als 20 Fällen, die nicht in die Berechnung des Perzentilwertes einfließen. Sind in dieser Gruppe mehr als 5 % der Standorte oberhalb des berechneten Perzentilwertes, so können auch insgesamt mehr als 5 % der Standorte rechnerisch auffällig werden. In den vorliegenden Extremfällen (QI 54120 und QI 54017) werden sogar mehr als doppelt so viele Standorte auffällig, wie nominell durch den Perzentilreferenzwert angestrebt. Abbildung 3 illustriert diesen Sachverhalt für den Qualitätsindikator mit QI-ID 54017.

[16] vgl. *RAW-10*- bzw. *RAW-20*-Methode in (Paddock 2014)

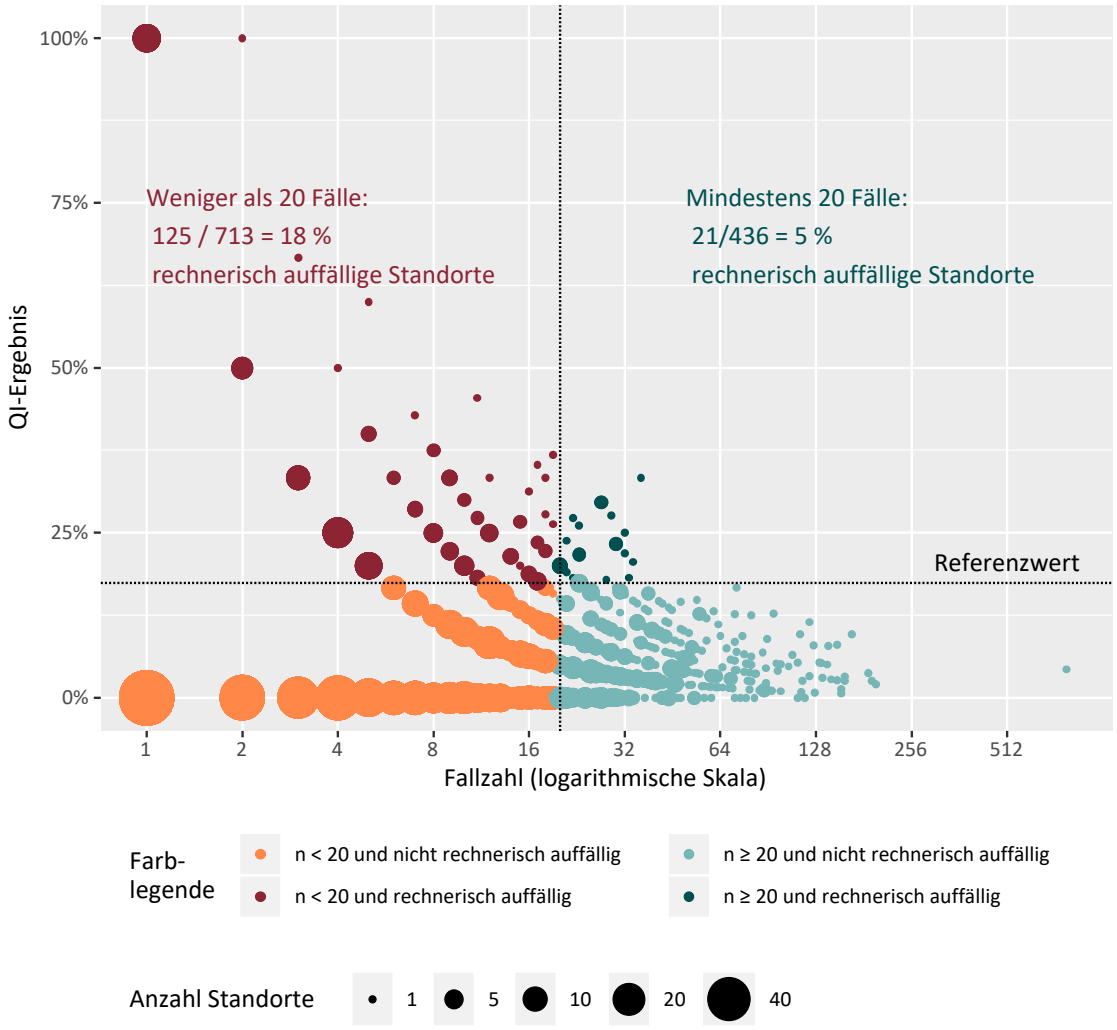

Abbildung 3: Standortergebnisse im Qualitätsindikator QI-ID 54017 „Allgemeine Komplikationen bei Hüftendoprothesen-Wechsel bzw. -Komponentenwechsel“ differenziert nach Fallzahl.

Gezeigt werden Standortergebnisse differenziert nach Fallzahl. Zur besseren Darstellung der Fallzahl wurde eine logarithmische Achsenskalierung zur Basis 2 gewählt, d. h. der Abstand zwischen zwei vertikalen Linien auf der x-Achse entspricht immer einer Verdopplung der Fallzahl. Die Größe der Punkte ist proportional zur Anzahl an Standorten mit gleicher Fallzahl und Ergebnis. Horizontal eingezeichnet ist der Referenzwert, der als 95. Perzentil aller Standorte mit mindestens 20 Fällen in der Grundgesamtheit des Indikators berechnet wird. 146 von 1.150 Standorten (12,7 %) wurden in diesem Indikator als rechnerisch auffällig eingestuft. Typisch für diese Funnelplot-Darstellung sind die erkennbaren Hyperbeläste 1/n, 2/n, 3/n etc. entlang derer die Standortergebnisse mit steigender Fallzahl n verlaufen. Die Grafik ist in vier Farbsegmente nach rechnerischer Auffälligkeit (ja/nein) und Fallzahlgruppe (< 20, ≥ 20) unterteilt. Während der Referenzbereich so bestimmt wird, dass genau 5 % der Standorte mit mindestens 20 Fällen auffällig werden, führt die Anwendung des gleichen Referenzbereiches auf die Standorte mit weniger als 20 Fällen dazu, dass in dieser Fallzahlgruppe ein deutlich über 5 % liegender Teil von 18 % der Standorte als rechnerisch auffällig eingestuft wird.

2.6.3 Qualitative Auffälligkeiten je Qualitätsindikator

Charakterisiert man den Aufwand des Strukturierten Dialoges anhand der absoluten Zahl rechnerischer Auffälligkeiten, so stellt sich die Frage wodurch bzw. ob ein geeignetes Maß für den „Nutzen“ des Strukturierten Dialoges im Sinne einer Aufwand-Nutzen-Gegenüberstellung definiert werden kann. In Abschnitt 5.2 werden Ansätze einer strukturierten Aufwand-Nutzen-Betrachtung bzw. Verlust-Betrachtung für den Strukturierten Dialog vorgestellt. Im Folgenden wird der Anteil an durch den Strukturierten Dialog identifizierten Qualitätsdefiziten, bezogen auf alle rechnerischen Auffälligkeiten bzw. alle Stellungnahmen, als eindimensionales Maß für den „Nutzen“ des Strukturierten Dialoges dargestellt. Dies ermöglicht eine Betrachtung von Effizienz im Sinne einer Aufwand-Nutzen-Abwägung, klammert allerdings aus, dass nicht identifizierte Qualitätsdefizite „immaterielle Kosten“ des Gesundheitssystems sind, die im Strukturierten Dialog nicht messbar sind. Eine wichtige Einschränkung der folgenden Analyse ist, dass die qualitativen Bewertungen des Strukturierten Dialogs bezogen auf die betrachteten rechnerischen Auffälligkeiten i. d. R. keine Sensitivität und Spezifität von 100 % aufweisen, d. h. es ist nicht sicher, dass Qualitätsdefizite nach einer rechnerischen Auffälligkeit immer entdeckt werden und nicht jede Bewertung als qualitativ auffällig identifiziert sicher ein Qualitätsdefizit.[17] Die qualitativen Bewertungen des Strukturierten Dialogs sind somit kein Goldstandard, der sich zum Benchmarking verschiedener statistischer Methoden zur quantitativen Auffälligkeitseinstufung eignet.

Abbildung 4 (oben) zeigt für jeden Qualitätsindikator die Anzahl an rechnerischen Auffälligkeiten (x-Achse) und den Anteil an qualitativen Auffälligkeiten unter allen rechnerisch auffälligen Standorten. Da nicht jeder rechnerischen Auffälligkeit im Strukturierten Dialog nachgegangen wird, wird in Abbildung 4 (unten) der Anteil an qualitativen Auffälligkeiten auf die Zahl an angeforderten Stellungnahmen bezogen.

[17] Andernfalls, bei perfekter Sensitivität und Spezifität, gäbe es auch keine Heterogenität der Bewertungen, wie sie in 2.5 deutlich nachgewiesen ist.

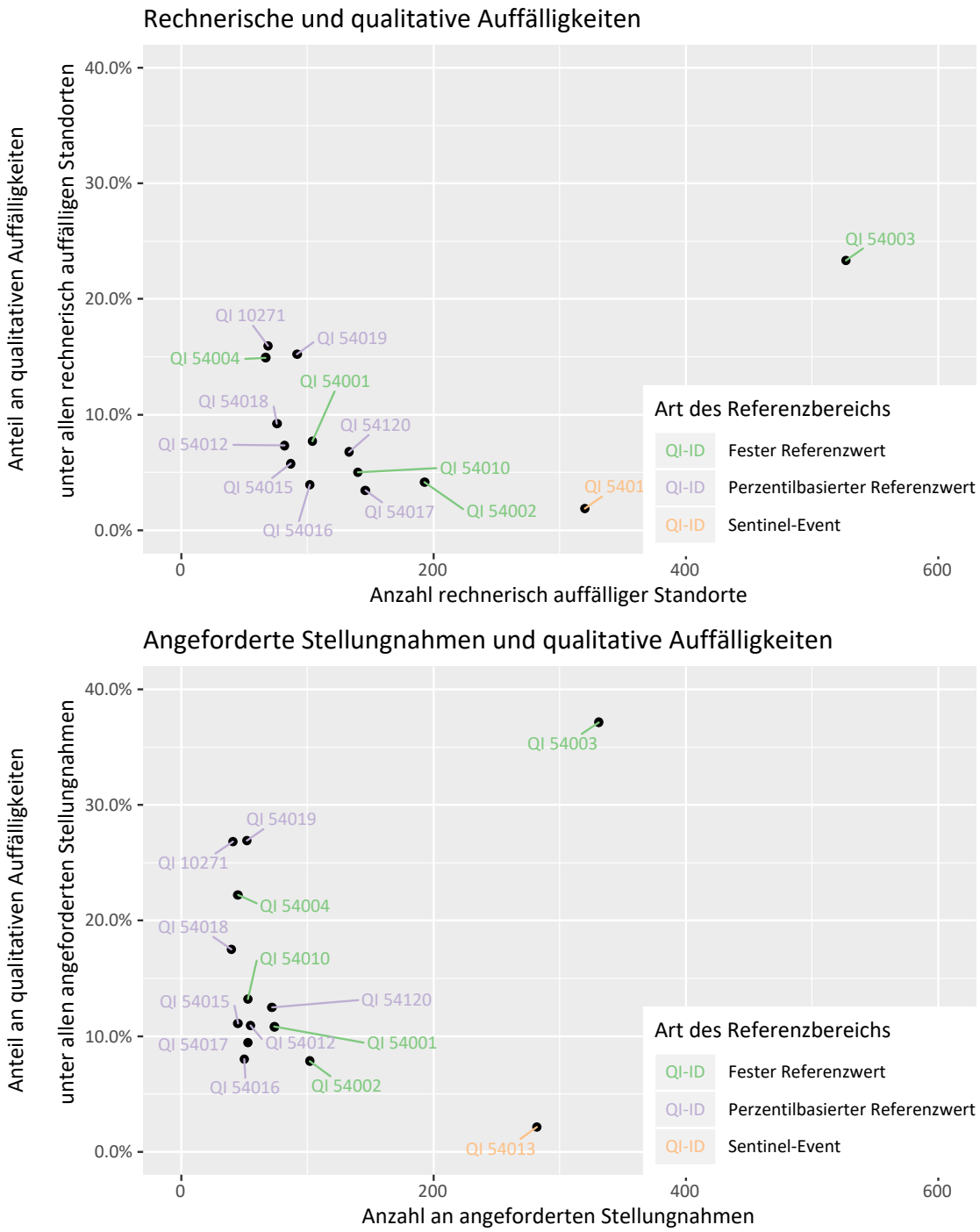

Abbildung 4: Qualitative Auffälligkeiten in Relation zu rechnerischen Auffälligkeiten (oben) und angeforderten Stellungnahmen (unten). Daten des Erfassungsjahres 2017.

Es fällt auf, dass der Anteil der qualitativen Auffälligkeiten an den rechnerischen Auffälligkeiten je nach Qualitätsindikator sehr unterschiedlich ist. Während eine rechnerische Auffälligkeit im Sentinel-Event-Indikator zur Sterblichkeit (QI-ID 54013) nur sehr selten (< 2 %) qualitative Auffälligkeiten nach sich zieht, wird im Indikator zur präoperativen Verweildauer (QI-ID 54003) mehr als jede vierte rechnerische Auffälligkeit auch als qualitativ auffällig bewertet. Die Mehrheit aller Indikatoren des HEP-Verfahrens erreicht weniger als 10 % qualitative Auffälligkeiten

unter allen rechnerischen Auffälligkeiten. Bezieht man die Zahl der qualitativen Auffälligkeiten auf die Anzahl an Stellungnahmen, zeigt sich für die meisten Qualitätsindikatoren eine höhere „Trefferquote“ im Sinne der obigen Auswertung. Insgesamt wurden bei ca. 61 % aller rechnerisch auffälligen Ergebnisse Stellungnahmen vom Leistungserbringer angefordert. Betrachtete man die qualitative Einstufung des Strukturierten Dialoges im Sinne eines diagnostischen Tests als „Goldstandard“[18], an der die Güte der rechnerischen Auffälligkeitseinstufung zu messen wäre, dann gäbe es im HEP-Verfahren über alle 14 Indikatoren berechnet ca. 90 % falsch positive rechnerische Auffälligkeiten. Ein Grund für diese hohe Zahl falsch positiver Ergebnisse ist möglicherweise darin zu sehen, dass das HEP Verfahren aus sehr vielen Ergebnisindikatoren besteht. Angesichts zum Teil nicht-risikoadjustierter Ergebnisindikatoren bzw. Risikoadjustierungen bei denen wichtige aus der Literatur bekannte Risikofaktoren nicht eingehen können, weil sie nicht erhoben werden, ist die Bewertung der Ergebnisqualität z. B. bei Todesfällen und Komplikationen eine besondere Herausforderung für den Strukturierten Dialog. Ein Grund dafür ist, dass strukturelle Probleme beim Leistungserbringer in diesem Fall deutlich schwieriger zu erkennen sind als beispielsweise bei Prozessindikatoren – besonders bei kleinen Fallzahlen.

Ein weiterer Grund für falsch positive Bewertungen ist auch in mangelnder Dokumentations- und Datenfeldqualität zu sehen. Beispielsweise wurde unter allen in HEP bewerteten Stellungnahmen ca. 13 % Dokumentationsprobleme im Sinne der Bewertungskategorien „D50“, „D51“ und „D99“ festgestellt. Angesichts des absoluten Aufwands, den das HEP-Verfahren durch die hohe Anzahl rechnerisch auffälliger Ergebnisse für den Strukturierten Dialog verursacht, ist ein Anteil von 90 % falsch positiver Auffälligkeiten als sehr hoch anzusehen. Es stellt sich die Frage, ob jeder auf Landesebene beauftragten Stelle für die gründliche Bearbeitung und Bewertung dieser Vielzahl von Stellungnahmen ausreichende Ressourcen zur Verfügung stehen. Gegebenenfalls könnte ein Mangel an für den Strukturierten Dialog zur Verfügung stehenden Ressourcen auch eine Erklärung für hohe falsch-positive Bewertungen sein. Weiterhin ist anzuführen, dass in der aktuellen Form des Strukturierten Dialogs die Beweislast bei den Bewertungsstellen liegt, ein bestehendes Qualitätsdefizit anhand der Stellungnahme des Leistungserbringers genau benennen zu können, was wiederum eine weitere Erklärung für den hohen Anteil rechnerischen Auffälligkeitseinstufungen sein kann, die dann im Strukturierten Dialog als qualitativ unauffällig bewertet werden. Die Ursachen für hohe falsch-positive Zahlen sind letztlich nicht klar und eindeutig: Sensitivität und Spezifität des Strukturierte Dialogs, Sensitivität und Spezifität des Einstufungsverfahrens zur rechnerischen Auffälligkeit, Datenqualität, aber auch Inhaltsvalidität und Operationalisierung jedes einzelnen Qualitätsindikators haben Einfluss auf die Effizienz des QS-Verfahrens insgesamt.

[18] Dabei sind die in Abschnitt 2.5 benannten Einschränkungen hinsichtlich der bewertungsstellenabhängigen Heterogenität der qualitativen Bewertungen zu berücksichtigen.

2.6.4 Zusammenhang zwischen Fallzahl und Auffälligkeitseinstufung

Um das Zustandekommen der sehr hohen Anzahl rechnerischer Auffälligkeiten, die als qualitativ unauffällig bewertet werden, zu untersuchen, wird im Folgenden eine nach Fallzahl differenzierte Analyse durchgeführt. Damit soll gezeigt werden, dass es im gegenwärtigen Einstufungsalgorithmus fallzahlabhängige Verzerrungen gibt, die zu überdurchschnittlich hohen Raten rechnerisch auffälliger Leistungserbringer in bestimmten Fallzahlgruppen führen.

Eine Sonderrolle spielt dabei der Sentinel-Event-Indikator mit der QI-ID 54013 (Todesfälle während des akut-stationären Aufenthaltes bei geringer Sterbewahrscheinlichkeit). Da die Fallzahl in die rechnerische Auffälligkeitseinstufung nicht mit einfließt, wie es z. B. bei der Berechnung einer Rate der Fall ist, sind auch die Zusammenhänge zwischen Fallzahl und Auffälligkeitseinstufung andere, als z. B. bei Raten- oder O/E-Indikatoren. Wegen dieser Sonderkonstruktion wird der Qualitätsindikator im Folgenden einzeln analysiert, während alle anderen Qualitätsindikatoren nur zusammengefasst dargestellt werden. Der Indikator ist nicht als Beispiel für den gesamten Leistungsbereich zu verstehen.

Die folgenden Abbildungen zeigen jeweils Anzahl und Anteile an Indikatorergebnissen in den folgenden vier sich ausschließenden Kategorien:

- „nicht rechnerisch auffällig“: Indikatorergebnisse, die als nicht rechnerisch auffällig bewertet wurden.
- „rechnerisch auffällig und keine Stellungnahme angefordert“: Indikatorergebnisse, die als rechnerisch auffällig bewertet wurden, zu welchen allerdings keine Stellungnahme angefordert wurde und daher auch keine Qualitätsbewertung erfolgte
- „rechnerisch auffällig und Stellungnahme angefordert“: Indikatorergebnisse, die als rechnerisch auffällig bewertet wurden, zu welchen eine Stellungnahme angefordert wurde, die aber nicht als qualitativ auffällig im Sinne der Bewertungen „A40“, „A41“, „A42“, „A99“ eingestuft wurden.
- „qualitativ auffällig“: Indikatorergebnisse, die als rechnerisch auffällig bewertet wurden, zu welchen eine Stellungnahme angefordert wurde, und die als qualitativ auffällig im Sinne der Bewertungen „A40“, „A41“, „A42“, „A99“ eingestuft wurden.

Diese vier Kategorien werden farblich kodiert und in Dezilen der Standortfallzahlen getrennt dargestellt. Durch die Kategorisierung nach Dezilen ist gesichert, dass in jeder der unten abgebildeten Kategorie gleich viele Leistungserbringer enthalten sind, und somit die berechneten Anteile vergleichbarer statistischer Unsicherheit unterliegen.

Zunächst werden der Zusammenhang zwischen Fallzahl und rechnerischer Auffälligkeitseinstufung und der Zusammenhang zwischen Fallzahl und qualitativer Bewertung deskriptiv beschrieben. Anschließend werden in Abschnitt 2.6.4.3 mögliche Ursachen diskutiert. Bei der Interpretation der dargestellten Zusammenhänge zwischen Fallzahl und rechnerischer Auffälligkeit, zwischen Fallzahl und Einholung einer Stellungnahme, sowie zwischen Fallzahl und qualitativer Bewertung ist zu beachten, dass sich systematische Eigenschaften des Einstufungsverfahrens und reale Zusammenhänge zwischen Fallzahl und Ergebnisqualität anhand dieser Zahlen nicht trennen lassen. Auf diesen Aspekt wird ebenfalls im Abschnitt 2.6.4.3 eingegangen.

2.6.4.1 Zusammenhang zwischen Fallzahl und rechnerischer Auffälligkeit

Nicht Sentinel-Event-QI

Im Folgenden stellen wir nach Fallzahl differenzierte Auswertungen für die 13 nicht-Sentinel-Event Indikatoren des HEP-Verfahrens dar. Dabei wird auf eine Einzeldarstellung jedes Indikators verzichtet und eine aggregierte Darstellung der Ergebnisse vorgenommen. Abbildung 5 zeigt die Häufigkeit rechnerischer Auffälligkeiten und in Folge dessen angeforderte Stellungnahmen aufgeteilt in die zehn Dezile der Fallzahlverteilung. Die obere Grafik zeigt die absoluten Zahlen an Indikatorergebnissen, die als rechnerisch auffällig eingestuft wurden und in Dunkelgrün die Anzahl solcher Auffälligkeiten, für die auch eine Stellungnahme angefordert wurde. Die rote Horizontale markiert dabei die durchschnittliche Anzahl an rechnerischen Auffälligkeiten pro Dezil. Man erkennt, dass überdurchschnittlich viele Leistungserbringer aus den ersten Dezilen als rechnerisch auffällig eingestuft wurden, während die Anzahl der rechnerischen Auffälligkeiten mit steigender Fallzahl sinkt. Die untere Grafik stellt den Anteil der rechnerischen Auffälligkeiten dar, für die auch eine Stellungnahme angefordert wurde. Der Anteil an Stellungnahmen an den rechnerischen Auffälligkeiten liegt im untersten Dezil bei 31 % und steigt bis zum 4. Dezil auf 71 % an. Dies kann in Teilen durch die in der QSKH-Richtlinie vorgesehene Anwendung der sog. 1-Fall-Regel erklärt werden, wonach bei rechnerisch auffälligen Ergebnissen mit nur einem auffälligen Zähler-Fall ein Hinweis versandt werden darf. Vom 5. bis 10. Dezil ist kein monotoner Zusammenhang im Anteil angeforderter Stellungnahmen zu beobachten. Im 8. Dezil liegt der Anteil bei 53 %, im 10. Dezil bei 70 %.

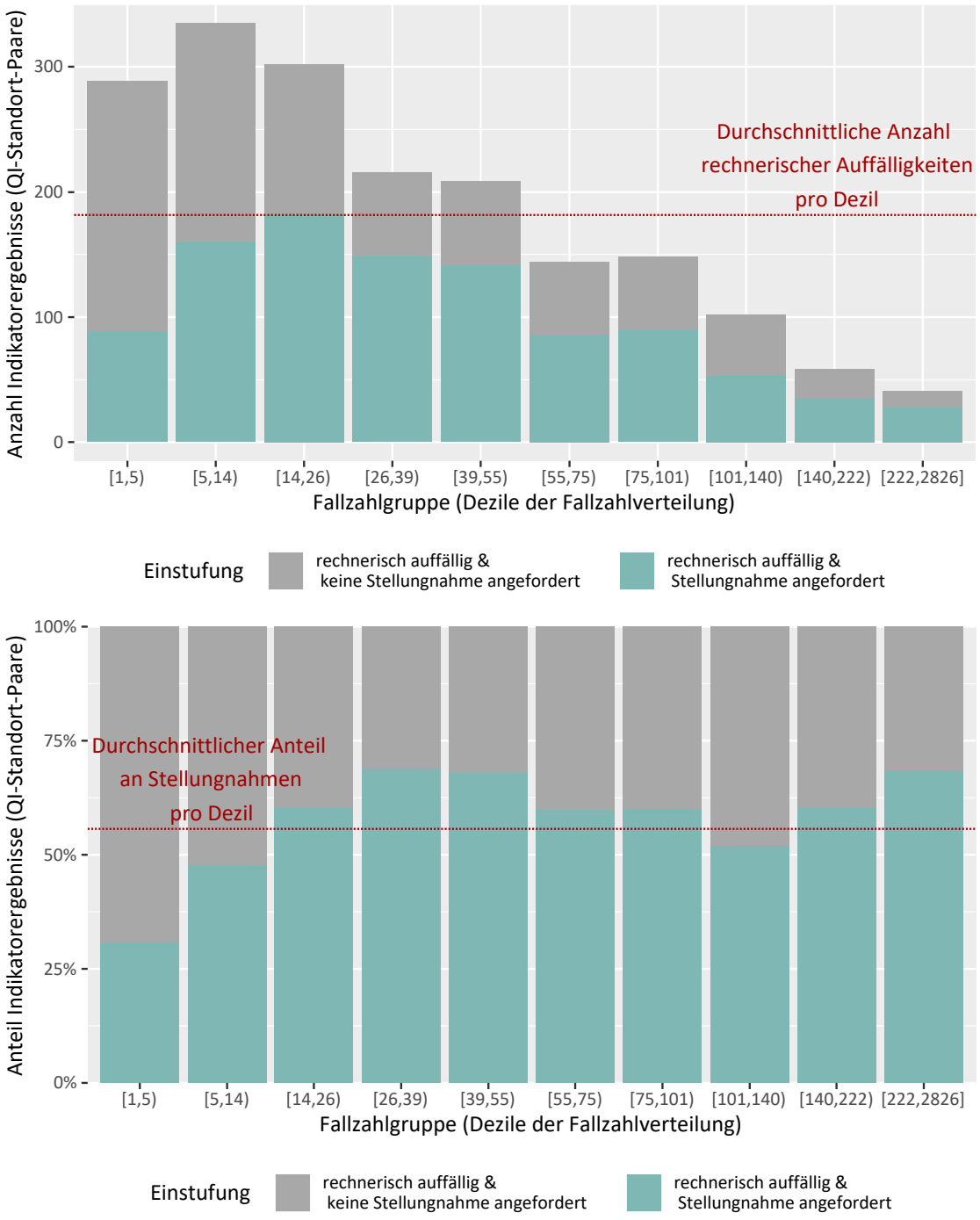

Abbildung 5: Rechnerische Auffälligkeitseinstufung und Stellungnahmen für Nicht-Sentinel-Event-Indikatoren im HEP-Verfahren differenziert in Dezilen der Fallzahlverteilung.

Sentinel-Event-QI

Abbildung 6 zeigt den Zusammenhang zwischen Fallzahl und rechnerischer Auffälligkeit beim Sentinel-Event-Indikator mit der QI-ID 54013 „Todesfälle während des akut-stationären Aufenthaltes bei geringer Sterbewahrscheinlichkeit“.

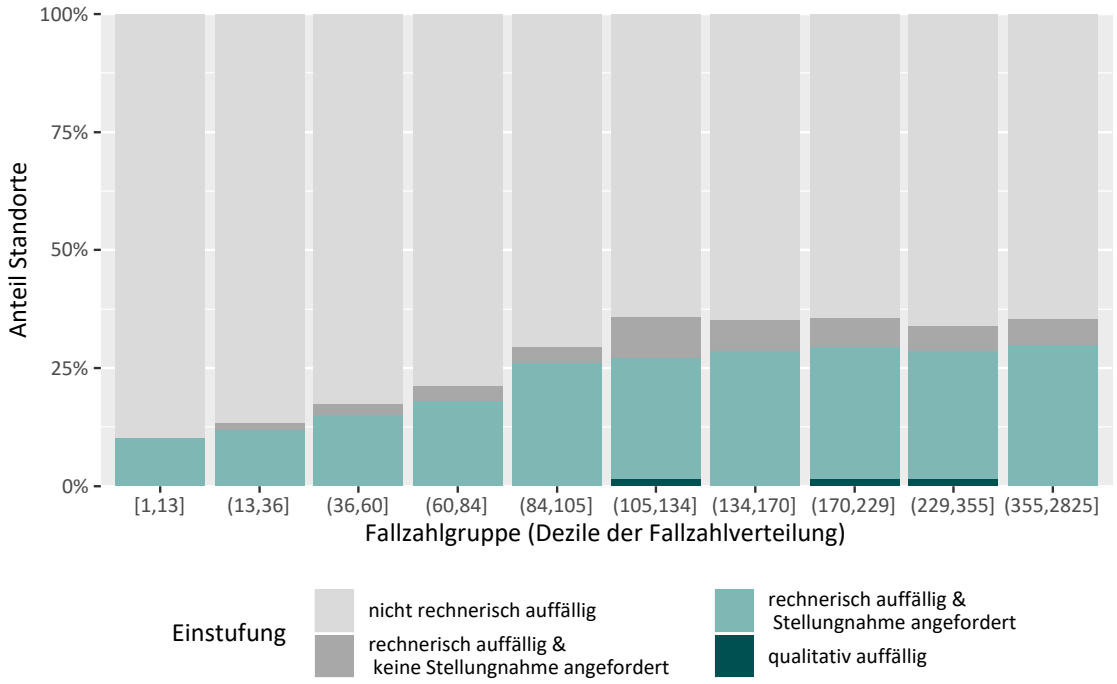

Abbildung 6: Rechnerische und qualitative Auffälligkeitseinstufung zum Sentinel-Event-Indikator „Todesfälle während des akut-stationären Aufenthaltes bei geringer Sterbewahrscheinlichkeit“ (QI-ID 54013) differenziert in Dezilen der Fallzahlverteilung.

Die Darstellung zeigt, dass der Anteil rechnerisch auffälliger Standorte vom 1. bis zum 6. Dezil der Fallzahlverteilung von 10 % auf 36 % steigt. Eine mögliche Erklärung für diesen Zusammenhang von Fallzahl und rechnerischer Auffälligkeit liegt in der speziellen Konstruktion des Qualitätsindikators. Als geringe Sterbewahrscheinlichkeit wird in der entsprechenden Datenbank der Qualitätsindikatoren ein Wert von höchstens 1,9 % definiert (IQTIG 2018e). Nur Fälle, deren Sterbewahrscheinlichkeit durch die Risikoadjustierung als geringer eingestuft wird, gehen in den Qualitätsindikator ein. Mit steigender Fallzahl steigt allerdings die Wahrscheinlichkeit, dass sich mindestens ein Todesfall pro Standort zu ereignet, auch wenn das Risiko für den einzelnen Fall als gering eingestuft wird.[19]

2.6.4.2 Zusammenhang zwischen Fallzahl und qualitativer Auffälligkeit

Nicht-Sentinel-Event-QI

Neben der Analyse des Zusammenhangs zwischen Fallzahlen und rechnerischen Auffälligkeiten ist auch eine nach Fallzahl differenzierte Auswertung der qualitativen Auffälligkeiten interessant. Abbildung 7 zeigt für die 13 Nicht-Sentinel-Event-Indikatoren den Anteil qualitativ auffälliger Standorte unter allen Standorten, von denen eine Stellungnahme angefordert wurde, und zwar aufgeteilt in den Dezilen der Fallzahlverteilung der QI-Standort-Paare. Zu erkennen ist, dass

[19] Nimmt man vereinfachend an, dass alle Fälle unabhängig voneinander mit einer „geringen“ Sterbewahrscheinlichkeit von π auftreten, so ist die Wahrscheinlichkeit eines Leistungserbringers mit n Fällen „zufällig auffällig“ zu werden durch $1-(1-\pi)^n$ gegeben, also eine mit n monoton wachsende Funktion. Mit π = 0,5 % und $n = 100$ ergäbe sich schon eine Wahrscheinlichkeit von ca. 39 %, ohne dass es Anhaltspunkte für ein besonderes Qualitätsdefizit gäbe.

der Anteil qualitativ auffälliger Ergebnisse vom 1. bis zum 6. Dezil monoton von 4 % auf 37 % steigt. Im 7. und 8. Dezil bleibt der Anteil bei über 30 % und fällt im 9. und 10. Dezil auf unter 20 % ab. Während also rechnerische Auffälligkeitseinstufungen in den kleinen Fallzahlbereichen überproportional häufig sind, sind die qualitativen Auffälligkeiten im mittleren Fallzahlbereich besonders häufig.

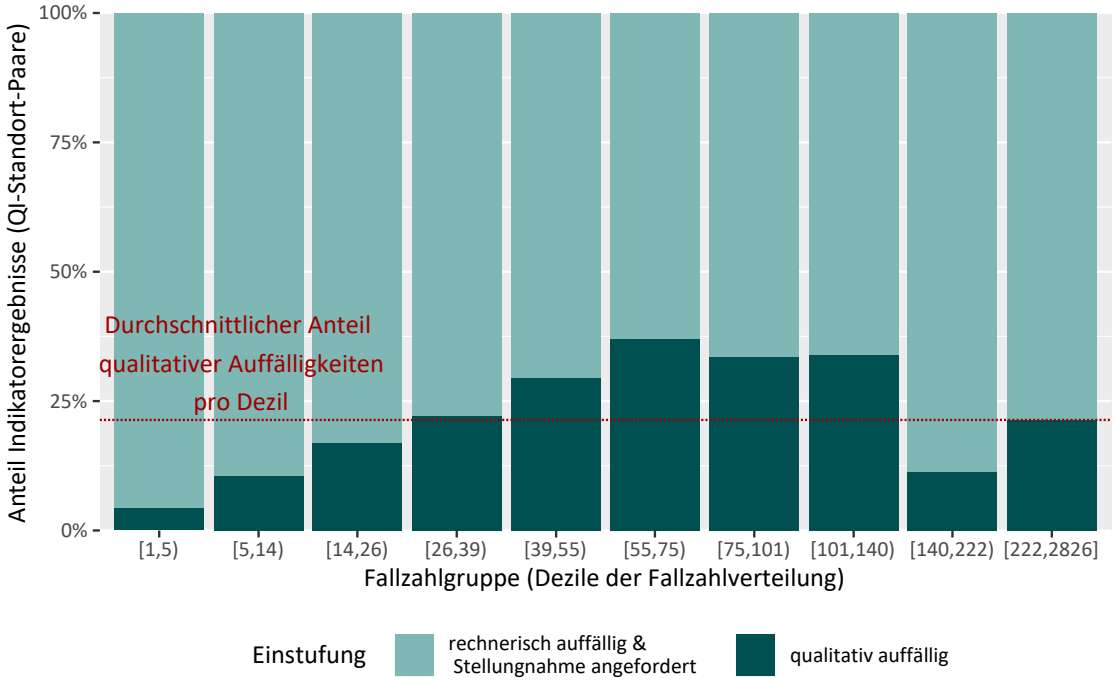

Abbildung 7: Anteil qualitativ auffälliger Standortergebnisse unter allen Ergebnissen, für die eine Stellungnahme angefordert wurde, differenziert nach Fallzahldezilen

Sentinel-Event-QI

Da es sich um einen Sentinel-Event-Indikator handelt, wird bei einem großen Teil (zwischen 84 % und 100 %) der rechnerischen Auffälligkeiten auch eine Stellungnahme angefordert. Qualitative Auffälligkeiten gibt es nur vereinzelt (vgl. Abbildung 6). Ein potentieller Zusammenhang zwischen qualitativer Auffälligkeit und Fallzahl ist wegen der Seltenheit der qualitativen Auffälligkeiten in diesem Indikator nicht zu bewerten.

2.6.4.3 Ursachen für den Zusammenhang zwischen Fallzahl und Auffälligkeitseinstufung

Die markanteste Beobachtung in Abbildung 5 zum Zusammenhang von Fallzahl und rechnerischer Auffälligkeitseinstufung bei Nicht-Sentinel-Event-Indikatoren ist die überproportionale Häufung von rechnerischen Auffälligkeiten bei kleiner Fallzahl. Wählt man wie in Abbildung 3 die Darstellung eines sog. Funnelplots für die nach Fallzahl differenzierte Darstellung der Standortergebnisse eines einzelnen Qualitätsindikators, so ist typischerweise eine breitere Streuung der Ergebnisse bei kleiner Fallzahl zu beobachten. Beispielsweise kann ein Leistungserbringer mit einer Fallzahl von fünf in einem ratenbasierten Indikator nur die Ergebniswerte 0 %, 20 %, 40 %, 60 %, 80 % und 100 % erreichen. Ein zusätzlicher Fall mit unerwünschtem Behandlungsergebnis (d. h. für den die Zählerbedingung des Indikators erfüllt ist) macht bei dieser Fallzahl somit immer einen Ergebnisunterschied von 20 % aus, während bei steigender Fallzahl der Beitrag

jedes Einzelfalls sinkt. Allerdings haben neben der Qualität der Versorgung durch einen Leistungserbringer auch andere, wie in Abschnitt 12.3 und Kapitel 15 von (IQTIG 2019a) beschriebene, Faktoren einen Einfluss auf das Indikatorergebnis eines Leistungserbringers. Diese werden im Folgenden unter „Stochastizität" bzw. „statistische Unsicherheit" subsumiert. Die Konsequenzen des Einflusses dieser stochastischen Komponente auf das beobachtete Indikatorergebnis ist für Leistungserbringer mit kleiner Fallzahl größer als für Leistungserbringer mit großer Fallzahl.

Prinzipiell wird die Möglichkeit weiterer Einflussfaktoren auf das Indikatorergebnis bereits durch die Wahl des Referenzbereiches berücksichtigt. Ein fester Referenzwert von 15 % im Indikator zur präoperativen Verweildauer sieht beispielsweise bereits vor, dass ein Leistungserbringer die Vorgaben zur präoperativen Verweildauer bei 15 % der Patientinnen und Patienten überschreiten darf. Es wird also angenommen, dass für bis zu 15 % der Patientinnen und Patienten Einflussfaktoren vorliegen, die die Überschreitung der präoperativen Verweildauer notwendig machen. Für einen Leistungserbringer mit weniger als sieben Fällen bedeutet der Referenzbereich dennoch, dass ein einziger Fall zur rechnerischen Auffälligkeit führen kann. Die gegenwärtige rechnerische Auffälligkeitseinstufung der QSKH-Richtlinie berücksichtigt nicht, dass die statistische Unsicherheit wesentlich von der Fallzahl abhängt. Es findet somit keine adäquate Berücksichtigung statistischer Unsicherheit statt,[20] was in Teilen die Häufung rechnerischer Auffälligkeiten bei kleinen Fallzahlen erklären kann. Tatsächlich kann man unter einfachen Modellannahmen zeigen, dass die Wahrscheinlichkeit trotz ausreichender Qualität rechnerisch auffällig zu werden (Fehler 1. Art) mit der Fallzahl des Standortes sinkt.[21]

Geht man davon aus, dass ein Grund für die Häufung rechnerischer Auffälligkeiten bei kleinen Fallzahlen im nicht Berücksichtigen statistischer Unsicherheit bei der rechnerischen Auffälligkeitseinstufung liegt, so wäre die Konsequenz eine höhere Zahl falsch positiver Ergebnisse bei kleinen Fallzahlen, was wiederum den beobachteten Zusammenhang zwischen Fallzahl und qualitativer Auffälligkeitseinstufung in Abbildung 7 teilweise erklären könnte.

Die bisherigen Darstellungen zur Fallzahlabhängigkeit der rechnerischen Auffälligkeitseinstufung können nicht unterscheiden, ob die beobachteten Tendenzen reale Zusammenhänge zwischen Fallzahl und Versorgungsqualität darstellen, oder ob es sich um Artefakte des Einstufungsalgorithmus bzw. des Einstufungsprozesses handelt. Die folgende Analyse zeigt, dass die beschriebenen Phänomene maßgeblich mit dem Einstufungsverfahren an sich zusammenhängen. Dafür wird exemplarisch der Qualitätsindikator „Allgemeine Komplikationen bei endoprothetischer Versorgung einer hüftgelenknahen Femurfraktur" (QI-ID 54015) betrachtet. Ziel der Analyse ist es, Differenzen zwischen dem Fallzahl-Ergebnis-Zusammenhang des Einstufungsverfahrens und dem Fallzahl-Ergebnis-Zusammenhang tatsächlicher Versorgungsqualität aufzudecken. Um dies zu erreichen, wird der Fallzahl-Ergebnis-Zusammenhang über alle Einzelfälle im Verfahren mit dem Fallzahl-Ergebnis-Zusammenhang der rechnerischen Auffälligkeitseinstufung

[20] Durch die in der §10(2) Satz 4 der QSKH-Richtlinie festgehaltene 1-Fall-Ausnahmeregel wird versucht dies auszugleichen. Die Richtlinie erlaubt in Fällen für die sich die rechnerische Auffälligkeit durch einen einzigen auffälligen Vorgang begründet, auf das Anfordern einer Stellungnahme zu verzichten.
[21] Vgl. Abschnitt 5.3.3

der Standorte gegenübergestellt. Dazu werden die gleichen Daten[22] auf unterschiedlichen Ebenen aggregiert. Durch eine Aggregation aller Einzelfälle, die in Standorten gleicher Fallzahl behandelt wurden, kann eine durchschnittliche Tendenz des tatsächlichen Zusammenhangs zwischen Fallzahl und Versorgungsqualität ermittelt werden. Aggregiert man die rechnerische Auffälligkeitseinstufung aller Standorte gleicher Fallzahl, so lässt sich der Zusammenhang zwischen Fallzahl und rechnerischer Auffälligkeitseinstufung ermitteln. Folgen die so ermittelten Fallzahl-Ergebnis-Zusammenhänge der beiden Aggregationsebenen nicht dem gleichen qualitativen Verlauf, so ist davon auszugehen, dass der Einstufungsalgorithmus systematische Verzerrungen des realen Zusammenhangs von Fallzahl und Versorgungsqualität bewirkt, welcher durch die Aggregation auf Bundesebene approximiert wird.

In Abbildung 8 werden oben die Ergebnisse des Qualitätsindikators auf Fallebene differenziert nach Fallzahl des behandelnden Standortes dargestellt. Grüne Punkte am oberen Rand der Grafik symbolisieren Behandlungsfälle, bei denen allgemeine Komplikationen während des stationären Aufenthaltes aufgetreten sind („Fälle mit interessierendem Ereignis“). Blaue Punkte am unteren Rand der Grafik stellen Vorgänge ohne Komplikationen dar. Die Größe der Punkte ist dabei proportional zur Anzahl an Behandlungsfälle je Fallzahl und Behandlungsergebnis. Durch die gelbe Linie wird dargestellt, wie hoch der modellbasiert geschätzte Anteil an Patientinnen und Patienten mit allgemeinen Komplikation ist, die in Standorten einer gegebenen Fallzahl behandelt wurden. Dieser Grafik liegt dabei ein additives logistisches Regressionsmodell zu Grunde, welches den Zusammenhang zwischen Fallzahl und dem Behandlungsergebnis durch einen Spline modellbasiert interpoliert und glättet (vgl. Wood (2006)). Man erkennt, dass der Anteil an Patientinnen und Patienten mit allgemeinen Komplikationen unabhängig von der Fallzahl bei einem gleichmäßigen Niveau von ca. 12 % liegt.

[22] QI-Ergebnisse aus dem Erfassungsjahr 2017 zum Indikator mit der QI-ID 54015

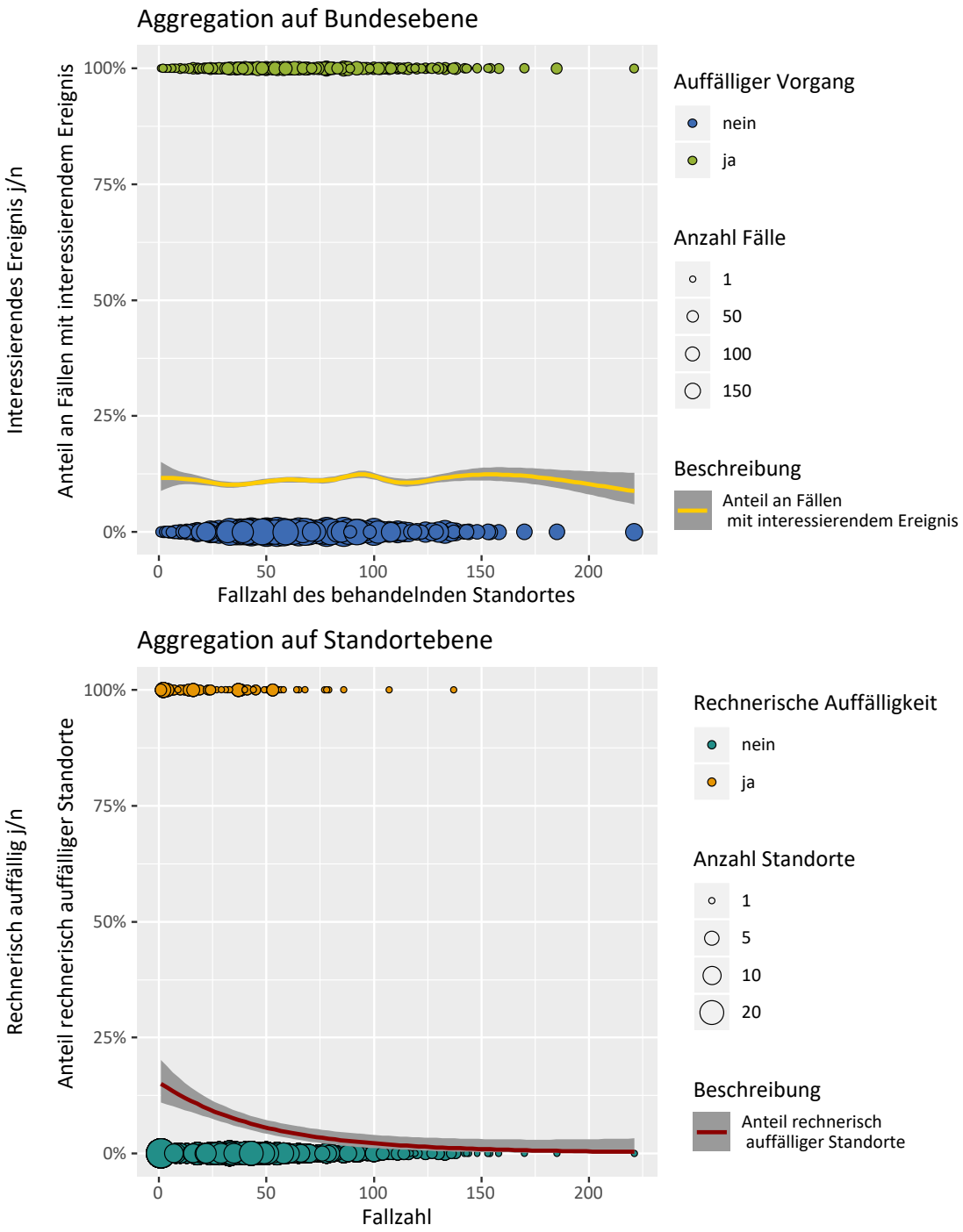

Abbildung 8: Fallzahlabhängigkeit der Behandlungsqualität auf Bundesebene (oben) und der rechnerischen Auffälligkeitseinstufung (unten) zum Qualitätsindikator „Allgemeine Komplikationen bei endoprothetischer Versorgung einer hüftgelenknahen Femurfraktur“

Demgegenüber steht die rechnerische Auffälligkeitseinstufung in Abbildung 8 unten. In dieser Grafik symbolisieren orange Punkte am oberen Rand Standorte, die als rechnerisch auffällig eingestuft wurden. Türkise Punkte am unteren Rand stellen Standorte dar, die nicht rechnerisch auffällig wurden. Die Größe der Punkte ist dabei proportional zur Anzahl an Standorten mit der

gleichen Fallzahl und rechnerischen Auffälligkeitseinstufung. Die rote Linie beschreibt den modellbasiert geschätzten Anteil rechnerisch auffälliger Standorte pro Fallzahl. Ähnlich wie in der oberen Grafik wird diese Linie durch ein additives logistisches Regressionsmodell geschätzt, welches den Zusammenhang zwischen Fallzahl und rechnerischer Auffälligkeitseinstufung durch einen Spline interpoliert und glättet. Zu erkennen ist, dass der Anteil rechnerisch auffälliger Standorte monoton mit steigender Fallzahl fällt. Dies steht im Widerspruch zum Fallzahl-Ergebnis-Zusammenhang auf Bundesebene, der in der oberen Grafik dargestellt ist. Dieses Phänomen wird von George et al. (2017) als „Mismatch of general and individual advice" bezeichnet. Die obere Grafik legt nahe, dass die Fallzahl bei der Wahl eines Krankenhauses irrelevant ist (*general advice*). Jedoch sind Standorte mit kleinerer Fallzahl häufiger rechnerisch auffällig als Standorte mit größerer Fallzahl, sodass die quantitative Bewertung des einzelnen Standortes bei kleinerer Fallzahl tendenziell schlechter ausfällt als für größere Standorte (*individual advice*). Theoretisch ließe sich dieser Widerspruch durch stark ausgeprägte Heteroskedastizität erklären, d. h. wenn ein starker Zusammenhang zwischen Fallzahl und *Heterogenität* der Versorgungsqualität vorläge, ohne dass die durchschnittliche Versorgungsqualität selbst fallzahlabhängig wäre. Beispielsweise könnte es hypothetische Konstellationen geben, in denen bei Standorten kleiner Fallzahl besonders viele Standorte mit überdurchschnittlicher und unterdurchschnittlicher Versorgungsqualität existieren, während Versorgungsunterschiede bei großer Fallzahl weniger ausgeprägt sind. In diesem Szenario wäre die mittlere Behandlungsqualität für alle Fallzahlen ähnlich, ein erhöhter Anteil auffälliger Standorte bei kleiner Fallzahl aber gerechtfertigt. Dies müsste sich aber auch in einer erhöhten Anzahl qualitativer Auffälligkeiten bei kleiner Fallzahl widerspiegeln, was sich aber nicht in den empirischen Ergebnissen des Strukturierten Dialogs zeigt (vgl. Abbildung 7).

Des Weiteren kann einfach gezeigt werden (vgl. Abschnitt 5.3.3), dass auch ohne Zusammenhang zwischen Fallzahl und Versorgungsqualität die Wahrscheinlichkeit rechnerisch auffällig zu werden mit der Fallzahl sinkt. Daher wird im Folgenden davon ausgegangen, dass der Zusammenhang zwischen Fallzahl und Auffälligkeitseinstufung auf ein Artefakt des Einstufungsalgorithmus zurück zu führen ist, der statistische Unsicherheit nicht berücksichtigt. Bei der Evaluation anderer möglicher Einstufungsverfahren ist zu fragen, ob die Eigenschaft, dass der Einstufungsalgorithmus qualitativ den gleichen Fallzahl-Ergebnis-Zusammenhang wiedergibt, der sich auf Fallebene nachweisen lässt („Fallzahlgerechtigkeit"), als Gütekriterium einzubeziehen ist.

2.6.5 Zwischenfazit

Der Aufwandsüberblick über die 14 Qualitätsindikatoren des HEP-Verfahrens macht deutlich, dass die Effizienzsteigerung im Strukturierten Dialog aus zweierlei Richtungen angegangen werden muss. Einerseits ist zu prüfen, bei welchen Qualitätsindikatoren es indikatorspezifischen Weiterentwicklungsbedarf gibt. Beispielsweise wird im Rahmen eines durch den G-BA erteilten Auftrags die Evidenzgrundlage des Qualitätsindikators zur präoperativen Verweildauer bereits geprüft (G-BA 2019). Auch hat sich im Strukturierten Dialog gezeigt, dass für den Qualitätsaspekt „Sterblichkeit in der Hüftendoprothesenversorgung" die angewandte Sentinel-Event-Methodik nicht zielführend ist, sodass die Bundesfachgruppe *Orthopädie und Unfallchirurgie* bereits eine

Änderung des Qualitätsindikators vorgeschlagen hat. Andererseits kann eine Effizienzsteigerung durch grundsätzliche methodische Weiterentwicklungen in der rechnerischen Auffälligkeitseinstufung erreicht werden. Hierzu haben die vorangegangenen Analysen gezeigt, dass die derzeitig verwendete Methode der rechnerischen Auffälligkeitseinstufung tendenziell zu viele Standorte mit kleiner Fallzahl als auffällig einstuft, weil statistische Unsicherheit in der Bestimmung der Auffälligkeitseinstufung nicht berücksichtigt wird. Dies führt zu einer hohen Zahl an rechnerischen Auffälligkeiten, die sich dann im Strukturierten Dialog nicht bestätigen – insbesondere bei kleiner Fallzahl.

Problematisch ist auch die Berechnung von perzentilbasierten Referenzwerten und die damit verbundene Auffälligkeitseinstufung, da die gegenwärtig verwendete Methodik dazu führt, dass insbesondere bei Qualitätsindikatoren mit einem hohen Anteil von Standorten mit einer Fallzahl kleiner 20 weit mehr Standorte auffällig werden, als durch das nominell angestrebte Niveau des Perzentilreferenzbereiches.

Grundsätzlich ist die absolute Zahl der im Strukturierten Dialog zu prüfenden Indikatorergebnisse und Stellungnahmen im Vergleich zur Anzahl der Krankenhausstandorte sehr hoch. In Abschnitt 2.5 dieses Berichtes wird thematisiert, dass die Landesstellen für Qualitätssicherung mit dieser Herausforderung sehr unterschiedlich umgehen, was sich beispielsweise darin ausdrückt, dass in einigen Bewertungsstellen häufiger auf das Anfordern einer Stellungnahme verzichtet wird und statt dessen nur ein Hinweis an rechnerisch auffällige Leistungserbringer versendet wird. Die Reduktion des Aufwands im Strukturierten Dialog kann unter diesem Gesichtspunkt auch einen wesentlichen Beitrag zur Reduktion der Heterogenität in den Bewertungen des Strukturierten Dialoges leisten. Gleichzeitig können mit der Reduktion des Aufwands mehr Ressourcen in die Durchführung jedes einzelnen Strukturierten Dialoges fließen.

3 Methodischer Hintergrund

Die Qualitätsmessung und -bewertung stellt ein zentrales Element der indikatorgestützten, gesetzlich verpflichtenden Qualitätssicherung dar. Nur auf Basis aussagekräftiger Qualitätsbewertungen sind die wichtigen weiterführenden Elemente, wie etwa qualitätsfördernde Maßnahmen oder die Veröffentlichung von Qualitätsergebnissen, sinnvoll. Ist die Aussagekraft der Qualitätsbewertung reduziert, werden darauf aufbauende Maßnahmen fehlgeleitet. Daher werden in diesem Kapitel die methodischen Kriterien für die Bewertung der Versorgungsqualität mittels Qualitätsindikatoren beschrieben sowie Kriterien für erfolgreiche Qualitätsförderung hergeleitet. Dazu werden zunächst einige Grundannahmen dargestellt.

3.1 Qualitätsmessung mittels Qualitätsindikatoren

Gemäß den Methodischen Grundlagen (IQTIG 2019a, Kap. 5) unterscheidet das IQTIG grundsätzlich zwischen Methoden der indikatorbasierten Messung der Versorgungsqualität und qualitativen Verfahren der Qualitätsmessung. Qualitätsindikatoren werden dabei als quantitative Größen verstanden, mittels derer Merkmale der Versorgung gemessen und dargestellt werden (IQTIG 2019a: 42). Sie erlauben eine statistische Aussage über die Erfüllung der Qualitätsanforderungen auf Aggregatebene in Bezug auf diese Merkmale, indem die Messergebnisse des Indikators mit einem Referenzbereich, der die Qualitätsanforderungen repräsentiert, verglichen werden (Soll-Ist-Vergleich[23]). Im Rahmen der Verfahren der datengestützten Qualitätssicherung und unabhängig von ihrer Richtliniengrundlage (QSKH-RL, Qesü-RL und DeQS-RL) repräsentiert die Referenzbereichsgrenze eines Indikators die von jedem Leistungserbringer erwartbare Versorgungsqualität – sie stellt Anforderungen dar, deren Erfüllung von jedem Leistungserbringer erwartet wird. Sofern ein Qualitätsindikator das Qualitätsmerkmal korrekt abbildet, bedeutet die Feststellung einer Abweichung vom Soll daher, dass das geforderte Qualitätsniveau der Versorgung nicht erreicht wurde.

Die indikatorbasierte Messung der Versorgungsqualität hat den Vorteil hoher Objektivität und Effizienz, da auch große Datenmengen anhand einheitlicher Regeln automatisiert aggregiert und ausgewertet werden können. Dementsprechend eignet sich diese Methode der Qualitätsmessung insbesondere für Situationen, in denen die Qualität der Versorgung bei einer Vielzahl von Leistungserbringern nach einheitlichen Bewertungskriterien beurteilt werden soll. Alternativ können auch qualitative Verfahren wie die des Peer-Review, Visitationen oder Audits für die Messung der Versorgungsqualität herangezogen werden. Diese Methoden basieren häufig auf Detailanalysen einzelner Behandlungsfälle (kasuistische Analyse), beispielsweise mittels Dokumentenanalysen (Begutachtung von Patientenakten), Gruppeninterviews (Dialogverfahren) oder Kombinationen davon (z. B. bei Peer-Review-Verfahren). Ein Vorteil solcher Verfahren ist, dass die besonderen Umstände des Einzelfalls berücksichtigt werden können, da eine potenziell

[23] Hier ist nicht der Soll-Ist-Abgleich im Rahmen der Soll-Ist-Statistik gemeint. Soll-Ist-Vergleich meint hier ganz allgemein den Vergleich zwischen dem geforderten und dem tatsächlichen Ergebnis des Qualitätsindikators.

größere Informationsgrundlage für die Beurteilung zur Verfügung steht als im Rahmen der indikatorbasierten Messung. Durch den größeren Aufwand, der bei diesen Verfahren aufgrund der weniger spezifischen Auswahl der zu betrachtenden Behandlungsfälle durch die geringeren Automatisierungsmöglichkeiten sowie die größere zu verarbeitende Informationsmenge resultiert, eignen sie sich eher in Fällen, in denen die Qualität weniger Leistungserbringer gemessen werden soll. Auch wenn die Ursachen für ein bestimmtes Indikatorergebnis oder wenn mögliche Maßnahmen zur Qualitätsverbesserung ermittelt werden sollen, eignen sich qualitative Bewertungsmethoden. Durch die geringeren Standardisierungsmöglichkeiten dieser Verfahren spielen allerdings implizite, subjektive Beurteilungskriterien der Durchführenden eine größere Rolle als bei der indikatorbasierten Qualitätsmessung (Donabedian 2003: 64 ff, Berwick 1991: 1218 f.), was zu einer niedrigeren Objektivität der Verfahren und damit eine zu einer Einschränkung der Vergleichbarkeit der Qualitätsmessungen unterschiedlicher Leistungserbringer z. B. für Auswahlentscheidungen.

Sollen die Ergebnisse von Qualitätsmessung auch im Kontext von *accountability* verwendet werden (vgl. Abschnitt 2.2) und Auswahlentscheidungen zwischen Leistungserbringer möglich sein, sind standardisierte Messmethoden wie die indikatorbasierte Qualitätsmessung besser geeignet. Steht ein *improvement*-Ansatz im Vordergrund der Verwendung der Qualitätsmessungen und stehen die Ressourcen für qualitative Verfahren zur Verfügung, beispielsweise, weil die Qualität nur weniger Leistungserbringer zu messen ist und sind Kriterien-basierte Auswahlentscheidungen von geringerer Bedeutung (bspw. in Notfallsituationen), sind qualitative Verfahren, wie Audits oder Peer-Reviews geeignete Methoden. Gemäß Beauftragung adressiert die in den folgenden Kapiteln beschriebenen Methodik die Qualitätsmessung mittels Qualitätsindikatoren.

3.2 Die methodische Funktion von Stellungnahmen

Die Ergebnisse von Messungen mit Qualitätsindikatoren werden nicht nur von der Versorgungsqualität der Leistungserbringer bestimmt, sondern auch von weiteren Einflussfaktoren, die nicht von den Leistungserbringern beeinflussbar sind, wie etwa Patientenrisikoprofile, stochastische Einflüsse oder im Indikator unberücksichtigte Versorgungssituationen (IQTIG 2019a: 181 f.). Um diese Einflussfaktoren auf das Indikatorergebnis bei der Bewertung der tatsächlichen Versorgungsqualität zu berücksichtigen, stehen unterschiedliche Maßnahmen zur Verfügung, die bei der Konstruktion eines Indikators, bei der Festlegung des quantitativen Bewertungskonzepts (Referenzbereich und statistisches Verfahren für Vergleiche mit diesem) oder bei der Anwendung der indikatorgestützten Bewertung auf den einzelnen Leistungserbringer angewandt werden können.

3.2.1 Einflussfaktoren auf das Indikatorergebnis

Bei der Konstruktion von Qualitätsindikatoren wird durch eine angemessene Festlegung der Berechnungsvorschrift und durch Maßnahmen der Risikoadjustierung der Einfluss bekannter, nicht qualitätsbezogener Faktoren auf das Indikatorergebnis so weit wie möglich reduziert. Für die Berücksichtigung von stochastischen Einflüssen auf die Indikatorergebnisse stehen beispielsweise verschiedene statistische Methoden zur Verfügung (siehe Abschnitt 5.1).

In der Praxis kann es aber selbst bei sorgfältiger Entwicklung eines Indikators entlang der Eignungskriterien (IQTIG 2019a: 135 ff. (Kapitel 10)) dazu kommen, dass das Indikatorergebnis bei manchen Leistungserbringern die Versorgungsqualität nicht angemessen abbildet. Dies kann durch besondere Konstellationen in der Versorgung geschehen, die in der Berechnungsvorschrift des Qualitätsindikators nicht berücksichtigt sind. Solche Konstellationen können zum einen dadurch zustandekommen, dass nicht alle Einflussfaktoren, die zu systematischen Abweichungen vom Qualitätsziel führen können, schon bei der Entwicklung des Indikators bekannt sind. Beispielsweise könnte erst im Routinebetrieb auffallen, dass bestimmte Begleiterkrankungen der Patientinnen und Patienten systematisch zu schlechteren Ergebnissen führen. Zum anderen können auch bekannte Einflussfaktoren im Qualitätsindikator nicht berücksichtigt sein, da ihre Ausprägung nicht für den Indikator erhoben wird. Ein Grund dafür, diese Informationen nicht zu erheben, kann z. B. der hohe Aufwand sein, der bei Erhebung sämtlicher bekannter, auch seltener, Risikofaktoren entstünde, und der dem Gebot der Praktikabilität und Datensparsamkeit entgegenstünde (vgl. Abschnitt 2.1). Auch eine unzureichende Objektivität und Reliabilität bestimmter Daten kann dazu führen, dass auf die Erhebung dieser Daten verzichtet wird.

Bei der Festlegung des Referenzbereichs eines Qualitätsindikators werden diese unbekannten oder nicht gemessenen Einflussfaktoren, die nicht vom Leistungserbringer zu verantworten sind, berücksichtigt. Dies geschieht dadurch, dass die Grenze eines Referenzbereichs nicht auf den Wert auf der Indikatorskala festlegt wird, der idealerweise erreicht würde, d. h. nicht auf 0 % oder 100 % (bei anteilsbasierten Indikatoren). Vielmehr sollte im Referenzbereich eine Abweichung von diesem Idealwert toleriert werden, die mindestens so groß ist wie das geschätzte Ausmaß, das die unbekannten und nicht gemessenen Einflussfaktoren im Mittel auf die Indikatorergebnisse haben.[24]

3.2.2 Prüfung der Validität je Messung

Treten unbekannte oder nicht gemessene Einflussfaktoren, die nicht in der Verantwortung der Leistungserbringer liegen, bei einem Leistungserbringer häufiger auf, als dies im Referenzbereich berücksichtigt ist, kann der Indikatorwert dieses Leistungserbringers den Referenzbereich verfehlen, obwohl die Versorgungsqualität zureichend ist. Es liegt dann eine Konstellation vor, in der der Abgleich zwischen Indikatorwert und Referenzbereich die Erfüllung der Qualitätsanforderungen für diesen Leistungserbringer nicht angemessen abbildet – die Validität des Indikators ist damit in diesem Fall in Zweifel gezogen. Um solche besonderen Konstellationen zu berücksichtigen, die nicht schon durch die Konstruktion des Indikators und sein Bewertungskonzept berücksichtigt sind, kann ein Stellungnahmeverfahren durchgeführt werden. Aus dieser Perspektive besteht die Funktion eines solchen Stellungnahmeverfahrens darin zu klären, ob das Verfehlen des Referenzbereichs durch einen Leistungserbringer nicht von diesem zu verantworten war. Diese Funktion eines Stellungnahmeverfahrens entspricht damit der Prüfung der Validität des Indikatorergebnisses im Einzelfall. Leistungserbringerübergreifend

[24] Zusätzlich wird meist auch toleriert, dass nicht immer optimale Qualität von den Leistungserbringern verlangt wird, sondern es werden nur von allen erwartbare Standards gefordert. Dies ist für die fachliche Bewertung von Bedeutung, siehe Abschnitt 6.2 (wird dort näher erläutert).

wird die Validität der Messung mittels Qualitätsindikatoren dadurch sichergestellt, dass ihre Entwicklung entsprechend den Eignungskriterien für Qualitätsmessung erfolgt (IQTIG 2019a: 135 ff. [Kapitel 10]). Das Stellungnahmeverfahren erfüllt damit die in Abschnitt 2.1 beschriebene Funktion, die Spezifität des Gesamtverfahrens (indikatorbasierte Messung einschließlich Prüfschritt 2) zur Feststellung von Qualitätsdefiziten durch Erhöhung der Validität zu steigern.

Aus dieser Perspektive ist es nicht Funktion des Stellungnahmeverfahrens, die Indikatorwerte für Leistungserbringervergleiche, zum Beispiel in Form von Rangreihenfolgen, zu korrigieren. Eine solche Korrektur der Indikatorwerte ist nicht sinnvoll, da dies nur für diejenigen Leistungserbringer möglich wäre, die eine Stellungnahme abgegeben haben. Damit ein fairer Vergleich zwischen Leistungserbringern anhand ihrer Indikatorwerte vorgenommen werden kann, müssten sowohl die Indikatorwerte der Leistungserbringer, die eine Stellungnahme abgegeben haben, als auch diejenigen, die keine Stellungnahme abgegeben haben, korrigiert werden. Zudem stehen in den QS-Verfahren nach QSKH-RL, DeQS-RL und plan. QI-RL Aussagen über die Erfüllung von Standards durch einen Leistungserbringer im Vordergrund.

Da es bei der Anwendung der Qualitätsindikatoren auf einzelne Leistungserbringer besondere Konstellationen geben kann, die in der standardisierten Qualitätsdarstellung anhand des Indikators nicht berücksichtigt sind, besteht die Funktion des Stellungnahmeverfahrens für die indikatorgestützte Qualitätsmessung darin, die Angemessenheit der Indikatoraussage für den jeweiligen Leistungserbringer zu prüfen. Die Prüfung der Validität der Messung für einen Leistungserbringer im Rahmen des Stellungnahmeverfahrens muss also die Frage beantworten, ob die Schlussfolgerung vom Indikatorergebnis auf die Erfüllung des Qualitätsmerkmals auch für die konkrete Konstellation dieses Leistungserbringers angemessen ist und sie alle wichtigen Einflussfaktoren, die nicht vom Leistungserbringer beeinflussbar sind, berücksichtigt. Ausgehend von dieser Funktion werden im Folgenden Kriterien für die Durchführung eines solchen Stellungnahmeverfahrens hergeleitet.

Stellungnahmen zu Indikatorergebnissen können prinzipiell und wurden vermutlich im bisherigen Strukturierten Dialog auch als Instrument der Qualitätsförderung verstanden. Durch die Aufforderung zur Stellungnahme zu einem „auffälligen“ Indikatorergebnis werden die Leistungserbringer aufgefordert, sich mit der Qualität ihrer Versorgung selbstkritisch auseinanderzusetzen. Im Sinne der *improvement* Strategie (Berwick et al. 2003) kann dies zur selbstständigen Identifikation von Qualitätsdefiziten führen und eine Umgestaltung der Versorgung durch die Leistungserbringer anregen, um Versorgungsqualität zu verbessern. Da jedoch die Indikatorergebnisse nicht nur für diesen Zweck, sondern auch im Rahmen von *accountability* verwendet werden, ist es wichtig, dass diese fair und aussagekräftig sind. Daher wird empfohlen, das Stellungnahmeverfahren als Schritt zur Prüfung der Validität der Messung des jeweiligen Indikatorergebnisses zu verstehen. Aus diesem Verständnis leiten sich die in Abschnitt 3.4 dargelegten Anforderungen an das Stellungnahmeverfahren ab.

3.3 Reichweite der Qualitätsaussage von Stellungnahmen

Gemäß der oben hergeleiteten Funktion von Stellungnahmen bezieht sich die fachliche Bewertung im Stellungnahmeverfahren auf das vom Indikator adressierte Qualitätsmerkmal. Es wird

nicht als Aufgabe des Stellungnahmeverfahrens verstanden, eine umfassende Qualitätsbewertung des gesamten Behandlungsprozesses vorzunehmen, zu dem das dem jeweiligen Indikator zugrundeliegende Qualitätsmerkmal gehört. So ist es nicht Ziel der fachlichen Bewertung, in den Behandlungsfällen, die zu einem auffälligen Indikatorergebnis geführt haben, grobe Behandlungsfehler zu finden, wenn diese keinen Bezug zum dargestellten Qualitätsmerkmal haben. Beispielsweise ist es im Rahmen der fachlichen Bewertung von Stellungnahmen zu einem Indikator zur Komplikationshäufigkeit nicht Ziel, unterlassene Aufklärungen in die Bewertung miteinzubeziehen, selbst wenn sie anhand einer Stellungnahme zu Tage treten. Dieses Vorgehen entspricht eher dem Verständnis von Qualitätsindikatoren als Aufgreifkriterien für einen nachfolgenden, gegebenenfalls umfassenden Qualitätsbewertungsprozess.

Für diese Beschränkung auf die jeweiligen Qualitätsmerkmale gibt es mehrere Gründe. Zum einen können Qualitätsmessungen Qualität nie vollumfänglich abbilden, sie können sich immer nur auf bestimmte Aspekte fokussieren. In der externen Qualitätssicherung sind dies die Qualitätsaspekte des Qualitätsmodells (Kap. 5.2 IQTIG 2019a: 44 ff.). Nach dem oben geschilderten Verständnis von Qualitätsindikatoren soll ein Qualitätsindikator zur Sterblichkeit bei akutem Herzinfarkt nicht die Qualität der Infarktbehandlung insgesamt beschreiben, sondern nur darstellen, ob die Sterblichkeit von Patientinnen und Patienten mit diesem Krankheitsbild bei einem Leistungserbringer zu hoch ist. Im Gegensatz zu Qualitätsindikatoren, für die das geforderte Qualitätsniveau über den Referenzbereich eindeutig definiert ist und durch Beschluss des G-BA normativen Charakter hat, liegen für die Gesamtqualität einer Behandlung keine standardisierten Beurteilungskriterien vor. Die Gesamtqualität der Behandlung korreliert auch nicht zwingend mit der Qualität in dem von einem einzelnen Indikator abgebildeten Merkmal. Vielmehr können verschiedene Aspekte des Behandlungsprozesses von einem Leistungserbringer mit unterschiedlicher Qualität durchgeführt werden (Shwartz et al. 2011). Die quantitative Auffälligkeit eines Indikatorergebnisses würde zudem bei einer über das Qualitätsmerkmal hinausgehenden Bewertung durch die Expertinnen und Experten möglicherweise als erhöhte A-priori-Wahrscheinlichkeit für eine unzureichende Gesamtqualität der Behandlung wahrgenommen und könnte daher zu einer Verzerrung bei der Beurteilung der Gesamtqualität führen.

Würden mit dem Stellungnahmeverfahren eine eigenständige Bewertung der Gesamtqualität angestrebt und würden Indikatoren lediglich als Aufgreifkriterium genutzt (siehe Abschnitt 2.2), wäre ein separates, von der indikatorbasierten Qualitätsmessung weitgehend unabhängiges Verfahren nötig (z. B. ein Auditverfahren). Die Bewertungskriterien, die für ein solches Verfahren benötigt würden, hätten nicht unbedingt einen direkten Bezug zu den konsentierten Anforderungen und Qualitätsmodellen der bisherigen indikatorbasierten QS-Verfahren. Die Qualitätsmessung mit Indikatoren ist zudem als ein quantitatives Verfahren bevorzugt für Themen mit nennenswerten Fallzahlen und Ereigniszahlen geeignet (IQTIG 2019a, S. 42 f.). Daraus folgt, dass die im Stellungnahmeverfahren von den Einrichtungen zur Verfügung gestellten Informationen über die Behandlung das Behandlungsgeschehen nicht vollumfänglich beschreiben können, sondern nur eine Auswahl an Informationen je geschilderten Behandlungsfall darstellen können. Diese Einschränkung der Informationsgrundlage steht einer angemessenen Gesamtbeurteilung eines Behandlungsfalls entgegen. Darüber hinaus müsste eine zuverlässige Gesamtaussage über die Behandlungsqualität bei einem Leistungserbringer alle Fälle des Leistungserbringers beim

interessierenden Behandlungsprozess oder wenigstens eine repräsentative Zufallsstichprobe dieser Fälle betrachten. Dieses Vorgehen würde im Vergleich zum bisherigen Vorgehen, bei dem nur die Fälle mit unerwünschtem Ereignis betrachtet werden (siehe Abschnitt 2.4), den Aufwand vervielfachen.

3.4 Gütekriterien für Bewertungsprozesse

Der Prozess der Bewertung in einem Stellungnahmeverfahren kann analog zum diagnostischen Prozess der Urteilsbildung verstanden werden, bei dem eine Vielzahl von Informationen von einem Experten, einer Expertin oder einem Expertengremium[25] berücksichtigt und zu einem Urteil verarbeitet werden (Schmidt-Atzert und Amelang 2012, Krohne und Hock 2015). Beispiele außerhalb des hier betrachteten Bewertungsprozesses des Stellungnahmeverfahrens sind die klinische Diagnosestellung oder die Entscheidung, ob ein Bewerber für eine Stelle geeignet ist. In diesen beiden Situationen fällen Expertinnen und Experten jeweils ein Urteil auf Basis einer gegebenen Informationsgrundlage und nach bestimmten, expliziten oder impliziten Entscheidungsregeln. Da die wissenschaftliche Literatur zur Urteilsbildung unter dem Begriff „klinische Urteilsbildung“ Fragestellungen untersucht hat, die mit dem Vorgehen der Qualitätsbewertung im Stellungnahmeverfahren verwandt sind, soll die Ähnlichkeit beider Sachverhalte genutzt werden, um auf Basis der Erkenntnisse aus der Urteilsforschung eine methodisch begründete Empfehlung für das Stellungnahmeverfahren zu entwickeln.

In der wissenschaftlichen Literatur zur Urteilsbildung wird zwischen sogenannter klinischer oder intuitiver Urteilsbildung und sogenannter statistischer oder „mechanischer“ Urteilsbildung unterschieden (Krohne und Hock 2015, Westen und Weinberger 2004). Im Gegensatz zur mechanischen Urteilsbildung, bei der Informationen nach formalen Regeln zu einem Urteil verarbeitet werden, integriert die Expertin oder der Experte bei der klinischen/intuitiven Urteilsbildung die Informationen in ihren/seinen Denkprozess. Hier zeigte die Urteilsforschung, dass die Güte der intuitiven Urteilsbildung, die auf impliziten Regeln beruht, der Güte der mechanischen Urteilsbildung auf Basis expliziter und konsistent angewandter Regeln unterlegen ist (Westen und Weinberger 2004, Krohne und Hock 2015, Dawes et al. 1989). Als Gründe werden einerseits die begrenzte und fehleranfällige menschliche Informationsverarbeitung und andererseits die flexible Anpassung von impliziten Regeln auf den Einzelfall angeführt. Es wird vermutet, dass sich bei der intuitiven Urteilsbildung individuelle Besonderheiten aufdrängen, die zu einem Abweichen von der (impliziten) Regel führen bzw. die dazu führen, dass die Auswahl der Regel von der Entscheidungssituation und von Eigenschaften des beurteilenden Individuums abhängt (Kruglanski und Gigerenzer 2011). Im Rahmen von beispielsweise Auswahl- und Zulassungsentscheidungen für Studienplatz- und Stellenbewerberinnen und -bewerber scheint diese intuitive Auswahl der Entscheidungsregel über viele Fälle hinweg zu ungünstigeren Ergebnissen zu führen als die mechanische Urteilsbildung (Kuncel et al. 2013).

[25] In der DeQS-RL wird der Begriff Fachkommission verwendet. Hier wird jedoch der allgemeinere Begriff „Expertengremium“ verwendet.

Bei der wissenschaftlichen Analyse des Prozesses der Urteilsbildung stellte sich die Unterscheidung zwischen der Art der zugrunde gelegten Informationen und der Art der Informationsverknüpfung als hilfreich heraus (Krohne und Hock 2015, Westen und Weinberger 2004). Durch explizite Vorgabe und konsistente Anwendung einer einheitlichen Entscheidungsregel auf der Basis einer standardisierten Informationsgrundlage kann die Güte von Entscheidungen erhöht werden (Westen und Weinberger 2004). Übertragen auf das Stellungnahmeverfahren bezieht sich die Art der zugrunde gelegten Informationen auf die Inhalte, die von den Expertinnen und Experten bei ihrer Qualitätsbewertung berücksichtigt werden dürfen. Die Art der Informationsverknüpfung bezieht sich auf die Regeln und Kriterien, nach denen die vorliegenden Informationen von dem Expertengremium zu einem Urteil (hinreichender vs. kein hinreichender Hinweis für unzureichende Qualität) verknüpft werden.

3.5 Ableitung von Anforderungen an das Stellungnahmeverfahren

Bislang existieren weder in der QSKH-RL noch in der DeQS-RL Regelungen, auf welche Art das Stellungnahmeverfahren (schriftliche Stellungnahme, Gespräch oder Begehung) durchzuführen ist und nach welchen Regeln die Bewertungsentscheidungen getroffen werden sollen. Abgeleitet aus der Forschung zur Urteilsbildung ist davon auszugehen, dass sowohl die Art des Stellungnahmeverfahrens (durch die unterschiedliche Informationsgrundlage) als auch die Abwesenheit expliziter Bewertungsregeln die Bewertung der Indikatorergebnisse beeinflusst. Um ein einheitliches Vorgehen bei der Qualitätsbewertung für alle Leistungserbringer auf der Grundlage gleicher Informationen zu gewährleisten, sollte daher eine Standardisierung des Stellungnahmeverfahrens erfolgen. Dies bedeutet, eine einheitliche Informationsgrundlage für die Bewertung der Stellungnahmen sicherzustellen. Des Weiteren sollten die Entscheidungsregeln für die Bewertung möglichst explizit sein, um zu verhindern, dass diese innerhalb und zwischen den urteilenden Personen bzw. Gruppen variieren.

Diese Anforderungen eines möglichst standardisierten, expliziten Vorgehens bei der Einholung, des Inhaltes und der Bewertungsregeln von und für Stellungnahmen lassen sich auch aus dem Gütekriterium der Objektivität ableiten. „Objektivität bedeutet, dass die Ergebnisse eines diagnostischen Verfahrens unabhängig davon zustande kommen, wer die Untersuchung, die Auswertung und die Interpretation durchführt“ (Schmidt-Atzert und Amelang 2012, S. 133). Sind Messergebnisse nach dieser Definition nicht objektiv, sind sie wenigstens zum Teil davon abhängig, wer die Messung, die Auswertung oder die Interpretation der Ergebnisse vornimmt. Objektivität ist darüber hinaus eine notwendige, aber nicht hinreichende Bedingung für die Reliabilität von Messergebnissen. Wird also (ausreichende) Reliabilität von dem Vorgehen im Strukturierten Dialog gefordert, muss zunächst eine ausreichende Objektivität sichergestellt werden.

Wird nun an der bisherigen Umsetzung des Strukturierten Dialogs eine starke Heterogenität bemängelt (siehe Beschluss zur Beauftragung sowie Abschnitt 2.5), bedeutet dies im Sinne dieser Definition, dass die Ergebnisse des Strukturierten Dialogs nicht objektiv genug sind. Dies bedeutet, dass neben der tatsächlichen Versorgungsqualität eines Leistungserbringers auch eine Rolle spielt, welche Stelle die Stellungnahme einholt, wie diese formuliert ist und welche Fachkommission diese bewertet. Diese Interpretation wird gestützt durch die Ergebnisse der empirischen

Analysen in Abschnitt 2.5, die zeigen, dass bewertungsstellenabhängige Heterogenität im Vorgehen vorliegt. Um die in der Beauftragung genannten Ziele einheitlicher, transparenter und nachvollziehbarer Qualitätsergebnisse zu erreichen, sollten demzufolge für die Prozessschritte der Qualitätsbewertung explizite Regeln vorgegeben werden (Standardisierung).

Es wird deutlich, dass ein explizites Vorgehen bei der Urteilsbildung wichtige Vorteile gegenüber einem impliziten Vorgehen hat und dass das explizite Vorgehen eine Voraussetzung für das in der Beauftragung angestrebte Ziel einer einheitlichen und nachvollziehbaren Bewertung ist.

3.6 Vor- und Nachteile explizierter Beurteilungsregeln im Stellungnahmeverfahren

Für das Stellungnahmeverfahren ist vor allem die Bewertung der zu einem Indikatorergebnis vorgebrachten Argumente von zentraler Bedeutung. Ein implizites Vorgehen würde bedeuten, dass die Argumente qualitativ durch die Expertinnen und Experten abgewogen würden. Für ein explizites Vorgehen würden vorab eine Rechenvorschrift und Gewichte für die Einflussfaktoren festgelegt werden, anhand derer eine Nachberechnung des Indikatorwerts für diese Einrichtung und ein erneuter Vergleich mit einem ggf. angepassten Referenzbereich durchgeführt würde. Expliztes und impliztes Vorgehen gehen mit unterschiedlichen Konsequenzen für die Umsetzung einher.

Vorteile expliziter Beurteilungsregeln bei der fachlichen Bewertung

Ein explizites Vorgehen bei der Bewertung der Einflussfaktoren in einem Stellungnahmeverfahren hat folgende Vorteile:

- Ein explizites Vorgehen einschließlich Nachberechnung des Indikatorergebnisses würde die Beurteilung als „Qualitätsdefizit“ bzw. „kein Hinweis auf ein Qualitätsdefizit“ stark standardisieren und damit die Objektivität des Bewertungsverfahrens steigern.
- Durch ein möglichst explizites Vorgehen kann ggf. die implizite Tendenz minimiert werden, in der fachlichen Beurteilung über das eigentlich zu beurteilende Qualitätsmerkmal hinauszugehen (siehe Abschnitt 3.3).
- Durch den Vergleich eines bereinigten Indikatorwerts mit einem (angepassten) Referenzbereich würde berücksichtigt, dass der Referenzbereich eines Indikators in der externen Qualitätssicherung durch Beschluss des G-BA normative Wirkung hat.
- Beim expliziten Vorgehen sind die Entscheidungsregeln transparent und ihr Einhalten prinzipiell überprüfbar (Dawes et al. 1989).
- Die Berechnung eines bereinigten Indikatorwerts ist auch mit statistischer Unsicherheit behaftet, die durch den Ausschluss von Behandlungsfällen aus der Grundgesamtheit im Vergleich zum ursprünglichen Indikatorwert noch erhöht sein kann. Diese Unsicherheit liegt impliziten Beurteilungen gleichermaßen zugrunde, würde aber durch ein explizites Vorgehen transparent gemacht.

Limitationen expliziter Beurteilungsregeln bei der fachlichen Bewertung

Gegen ein explizites Vorgehen in der Beurteilung im Rahmen eines Stellungnahmeverfahrens sprechen folgende Überlegungen:

- Durch explizite Regeln in der Beurteilung wird das Vorgehen deutlich komplexer: Wo vorher ein summarisch-implizites Urteil gefällt wurde, müssten dann mehrere Prozesse durchgeführt werden, wie eine explizite Bewertung jedes genannten Einflussfaktors, eine Nachberechnung des Indikatorwerts, ein Vergleich mit dem Referenzwert etc. Viele dieser Schritte können nicht ohne weitere technische Unterstützung durchgeführt werden.
- Für ein explizites Vorgehen müssten bereits A-priori-Bewertungsregeln für alle möglichen Einflussfaktoren vorliegen. Es ist aber charakteristisch für die zu bewertenden besonderen Konstellationen, dass eben nicht alle möglichen Einflussfaktoren vor Beginn des Stellungnahmeverfahrens bekannt sind.
- Es gibt Konstellationen, in denen die verfügbaren Informationen nicht ausreichen, um Entscheidungsregeln und Berechnungsvorschriften festzulegen oder diese sinnvoll anzuwenden.
- Eine Umstellung von einem eher impliziten auf ein expliziteres Vorgehen könnte für klinische Fachexpertinnen und Fachexperten, die berufsbedingt eher nach einem implizit-erfahrungsbasierten Kognitionsmuster vorgehen (Djulbegovic et al. 2014), ungewohnt sein.

Mit einer Steigerung der Objektivität des Stellungnahmeverfahrens geht auch eine gewisse Steigerung der Komplexität des Verfahrens einher. Je mehr Regeln explizit formuliert werden, desto mehr Regeln müssen auch umgesetzt werden. Dies weist auf einen Konflikt hin zwischen den in der Beauftragung genannten Zielen einer einheitlichen und transparenten (d. h. objektiven) Vorgehensweise einerseits und einem schlanken (d. h. aufwandsarmen) Verfahren andererseits hin. Eine aufwandsarme Qualitätsmessung kann im Allgemeinen vor allem mittels (automatisierbarer) quantitativer Methoden, wie der Indikator-basierten Qualitätsmessung im Vergleich zu qualitativen Verfahren, wie etwa Peer-Reviews oder Begehungen, erreicht werden (vgl. IQTIG 2019a, Kapitel 5.1)[26]. Sollen die Ergebnisse quantitativer Verfahren in einem zweiten qualitativen Prüfschritt analysiert werden, wird damit der Effizienzgewinn des quantitativen Verfahrens wenigstens teilweise aufgegeben. Werden an den zweiten qualitativen Prüfschritt auch noch ähnlich hohe Anforderungen an Objektivität angelegt wie an das quantitative Verfahren, erhöht sich der Aufwand zusätzlich. Außerdem setzt die qualitative Informationsgrundlage von Stellungnahmen Grenzen für die Standardisierung des Vorgehens. Aus dieser Gegenüberstellung wird deutlich, dass die methodischen Anforderungen, die sich aus dem beauftragten Ziel der Standardisierung ergeben, sowie das Ziel einer „Verschlankung des Verfahrens“ unter Beibehaltung eines qualitativen Prüfschritts gegeneinander abgewogen werden müssen. Dies erfolgt in Form von konkreten Empfehlungen in Kapitel 6.

[26] Neben dem hier gemeinten Aufwand für die Datenaus- und –bewertung fällt auch Aufwand für die Datenerhebung an. Für alle Daten, die nicht in Form von Sozialdaten, Patientenbefragungsdaten oder Daten, die automatisiert aus dem KIS oder der PVS auslesbar sind, fällt ein bedeutsamer Aufwand für die Erhebung bei den Leistungserbringern an.

3.7 Methodische Einordnung der Qualitätsförderung

Neben der Qualitätsbewertung und auf ihr aufbauend ist die Qualitätsförderung ein weiteres wichtiges Element des Strukturierten Dialogs. Wird ein Leistungserbringer nach der abschließenden Qualitätsbewertung als qualitativ auffällig eingestuft, ist das vorrangige Ziel der Qualitätsförderung, das einrichtungsinterne Qualitätsmanagement darin zu unterstützen, den Qualitätsmangel zu beheben.

Zu den Instrumenten externer Qualitätsförderung zählen neben dem Strukturierten Dialog klinische Audits, Visitationen, Audit-Feedback-Verfahren und Qualitätszirkel. Alle Instrumente ähneln dem Instrument des Peer-Review-Verfahrens. Auch der Strukturierte Dialog kommt dem Peer-Review am nächsten (Hensen 2016, Griem et al. 2013). Eine Evaluation des Strukturierten Dialogs wurde bislang nicht durchgeführt, sodass derzeit keine wissenschaftlichen Daten zu sogenannten „Erfolgsfaktoren“[27] für den Strukturierten Dialog vorliegen. Um dennoch eine methodische Einordnung der qualitätsfördernden Maßnahmen des Strukturierten Dialogs vornehmen zu können, wird auf die Literatur zu Erfolgsfaktoren für die Gruppe der Peer-Review-Verfahren zurückgegriffen (BÄK 2014, Chop und Eberlein-Gonska 2012, Gerlach 2001, Hudson et al. 2012, Rink 2013). Diese sollen hier herangezogen werden, um eine Einordnung des Strukturierten Dialogs im Hinblick auf den Erfolg der qualitätsfördernden Maßnahmen vorzunehmen und daraus Empfehlungen für die Weiterentwicklung abzuleiten. Bei dieser Einordnung und der Ableitung von Empfehlungen ist zu berücksichtigen, dass der Strukturierte Dialog durch die Richtlinienvorgaben einigen grundsätzlichen Regelungen unterlegen ist. Daher können manche Erfolgsfaktoren nur bedingt auf den Strukturierten Dialog übertragen werden. Die Erfolgsfaktoren beziehen sich nicht nur auf die Umsetzung der jeweils gezielten Maßnahme zur Behebung eines Qualitätsdefizits (kollegiales Gespräch oder Begehung), sondern auf alle Phasen des Peer-Review und betreffen damit alle Elemente des Strukturierten Dialogs. Im Folgenden werden Erfolgsfaktoren beschrieben, die durch Optimierung der Qualitätsbewertung eine Qualitätsförderung positiv beeinflussen.

In der Literatur zu Erfolgsfaktoren für Peer-Review-Verfahren werden die fachliche Kompetenz der Peers sowie die Zusammensetzung unabhängiger, interdisziplinärer Peer-Review-Teams als wichtige Erfolgsfaktoren für qualitätsfördernde Maßnahmen genannt (Eberlein-Gonska et al. 2017, Chop und Eberlein-Gonska 2012, Gerlach 2001, BÄK 2014, Rink 2013, Hudson et al. 2012). Um für die Qualitätssicherung ein Höchstmaß an Glaubwürdigkeit und Akzeptanz zu erreichen, ist die Glaubwürdigkeit und Kompetenz der Peers von Bedeutung (DeGEval 2008). Zudem ist davon auszugehen, dass durch eine entsprechende Expertise leichter eine Verknüpfung zwischen dem Einsatz der Qualitätsindikatoren und den Prozessen und Strukturen der Leistungserbringer hergestellt werden kann. Übertragen auf den Strukturierten Dialog und in Bezug auf seine Akzeptanz bei allen Beteiligten bedeutet das zum einen, dass die medizinische bzw. fach-

[27] Neben dem Begriff Erfolgsfaktor kommen in der Literatur auch die Begriffe Erfolgs- oder Qualitätskriterien vor. Diese Begriffe werden in der entsprechenden Literatur nicht trennscharf verwendet, ihnen allen ist jedoch gemein, dass es sich um Faktoren handelt , die einen positiven Einfluss auf das Gelingen qualitätsfördernder Maßnahmen haben.

liche Qualifikation und die praktische Erfahrung im jeweiligen Versorgungsbereich zentrale Auswahlkriterien für die Mitglieder der Fachkommission sein sollten. Zum anderen sollte die Zusammensetzung der Fachkommission alle wesentlichen, an der Versorgungsqualität beteiligten Berufsgruppen im zu bewertenden QS-Verfahren (Fachärztinnen und Fachärzte, Gesundheitsfachberufe und ggf. Hygieniker) involvieren, um eine interdisziplinäre Zusammensetzung zu gewährleisten. Da Interessenkonflikte das Urteilsvermögen verzerren können, kann es zu Fehleinschätzungen kommen (Lieb et al. 2011). Daher sollten für einen möglichst unabhängigen Bewertungsprozess bei der Wahl der Expertinnen und Experten deren Interessenkonflikte bekannt und geprüft worden sein.

Aus der Analyse der Ausgangssituation geht hervor, dass aktuell nicht einheitlich oder gar nicht geregelt ist, welche spezifischen Qualifikationen und wieviel Erfahrung notwendig sind, um als Expertin oder Experte ausgewählt zu werden. In Bezug auf die Zusammensetzung ist in der QSKH-RL festgelegt, dass die Fachkommissionsmitglieder von den jeweiligen Interessenvertretungen zu berufen sind. Durch diese Regelung alleine kann jedoch nicht von einer unabhängigen und interdisziplinären Zusammensetzung ausgegangen werden. Um den Erfolg und die Akzeptanz des Verfahrens durch eine möglichst objektive Bewertung der Versorgungsqualität zu steigern, sollten die genannten Aspekte bei dem Auswahlprozess der Mitglieder berücksichtigt werden. Zudem sollte für eine größere Akzeptanz des Verfahrens die interdisziplinäre Perspektive durch die Perspektive der Patientinnen und Patienten, die die eigentlichen Adressaten der externen Qualitätssicherung sind, ergänzt werden. Das bedeutet, dass auch Vertreterinnen und Vertreter der Patientinnen und Patienten als Mitglieder in das Expertengremium aufgenommen werden sollten.

Eine vorhandene *Bewertungssystematik* mit klaren Festlegungen z. B. zu den Bewertungsinhalten und zum Bewertungsvorgehen ist ein weiterer wesentlicher Erfolgsfaktor für qualitätsfördernde Maßnahmen (BÄK 2014, Chop und Eberlein-Gonska 2012, Rink 2013, Hudson et al. 2012). Welche methodischen Anforderungen an beide Faktoren zu stellen sind, wurde bereits in den vorherigen Abschnitten beschrieben. Zu ergänzen ist, dass die Bewertungssystematik einen großen Einfluss auf die Qualitätsförderung besitzt. Die Analyse der Ausgangssituation zeigt, dass bei der Umsetzung qualitätsfördernder Maßnahmen (wie z. B. dem kollegialen Gespräch) die Bewertung der Versorgungsqualität oft noch nicht abgeschlossen ist. Wurde noch keine Bewertung vorgenommen ist davon auszugehen, dass die offene Selbstreflexion des Leistungserbringers bezüglich bestehender Qualitätsdefizite unter dem Druck der noch ausstehenden Bewertung eingeschränkt sein könnte. Somit können qualitätsfördernde Maßnahmen nicht an den relevanten Problemen ansetzten und würden folglich fehlgeleitet. Für eine effiziente Qualitätsförderung ist demnach der Bewertungsvorgang abzuschließen, bevor qualitätsfördernde Maßnahmen eingeleitet werden.

Ein weiterer Erfolgsfaktor, der als Voraussetzung für den Erfolg von Peer-Reviews hervorgehoben wird, ist die *Sanktionsfreiheit* innerhalb des Verfahrens. In der Literatur wird beschrieben, dass durch ein *bedrohungsfreies Klima* die Bereitschaft der Betroffenen erhöht wird, die im Praxisalltag auftretenden Probleme offen anzusprechen und konstruktiv damit umzugehen. Zudem wird die Motivation, Veränderungen einzuleiten, erhöht (BÄK 2014, Hudson et al. 2012). Diese

Voraussetzung ist für den Strukturierten Dialog nicht gegeben, da die Bewertung eines Leistungserbringers bei Versorgungsmängeln eine Art *Sanktion* darstellt, insbesondere da qualitativ auffällige Ergebnisse für die Öffentlichkeit zugänglich und verpflichtend zu berichten sind. Zudem gibt die Richtlinie vor, dass bei einem qualitativ auffälligen Ergebnis weiterführende Maßnahmen einzuleiten sind. Dementsprechend ist der Strukturierte Dialog kein sanktionsfreies Verfahren. Einheitliche Regelungen, die die Bewertung der Versorgungsqualität strikt von der Qualitätsförderung trennen, würden jedoch eine freiere Qualitätsförderung im Anschluss an die Bewertung ermöglichen und könnten damit den Erfolg der Maßnahmen steigern.

Des Weiteren ist eine Regelung zur *Vertraulichkeit und zum Datenschutz* ein relevanter Erfolgsfaktor (BÄK 2014). Im Hinblick z. B. auf kollegiale Gespräche und Begehungen ist davon auszugehen, dass nur in einem vertraulichen Rahmen die Leistungserbringer eine selbstkritische Analyse vornehmen werden. Zudem wird auch durch die Einhaltung klarer Regelungen zum Datenschutz die Akzeptanz des Gesamtverfahrens gesteigert. Das führt dazu, dass qualitätsfördernde Maßnahmen gewinnbringender umgesetzt werden können. Klare und einheitliche Regelungen zum Datenschutz sind im Rahmen des Strukturierten Dialogs bereits für die Qualitätsbewertung anhand von Stellungnahmen von Bedeutung. In der Analyse der Ausgangssituation wurde beschrieben, dass im Hinblick auf die Qualitätsbewertung bei den direkten Verfahren der Datenschutz durch das Festlegen von formalen Kriterien (z. B. Wahrung der Anonymität der Patientinnen und Patienten und der Leistungserbringer) für die Stellungnahmen teilweise abgedeckt ist. Regeln im Umgang mit sensiblen Daten sollten in Abhängigkeit von den jeweiligen Prozessen des Strukturierten Dialogs festgelegt werden. Hier sollte zwischen der Bewertung der Versorgungsqualität anhand von Stellungnahmen einerseits und qualitätsfördernden Maßnahmen andererseits, die einen persönlichen Kontakt mit dem Leistungserbringer erfordern, unterschieden werden. Für beide Vorgänge ist die Verarbeitung unterschiedlicher Informationen (z. B. Stellungnahmen vs. Patientenakten, Operationsberichte, SOP des Leistungserbringers) notwendig. Des Weiteren ist das Setting, in welchem die Qualitätsbewertung und qualitätsfördernde Maßnahmen stattfinden, ein anderes. Aufgrund dieser Unterschiede ergeben sich für die Vertraulichkeit und den Datenschutz in Abhängigkeit davon, ob es sich um den Prozess der Qualitätsbewertung oder um qualitätsfördernde Maßnahmen handelt, unterschiedliche Anforderungen. Vor dem Hintergrund unterschiedlicher Anforderungen sollten folglich Vertraulichkeitserklärungen und Datenschutzregeln zwischen der Qualitätsbewertung und der Qualitätsförderung differenzieren.

Die *Freiwilligkeit der Teilnahme am Verfahren* wird in der Literatur als eine weitere Voraussetzung für den Erfolg qualitätsfördernder Maßnahmen beschrieben (BÄK 2014, Chop und Eberlein-Gonska 2012, Gerlach 2001, Nimptsch et al. 2016). Freiwilligkeit setzt die Überzeugung der Teilnehmer vom Nutzen des Verfahrens voraus, wodurch die Akzeptanz als auch die Veränderungsbereitschaft erhöht wird. Die Ziele des Verfahrens werden idealerweise die eigenen Ziele der teilnehmenden Krankenhäuser bzw. Leistungserbringer, sodass das Verfahren als selbstbestimmte Qualitätsentwicklung wahrgenommen wird (BÄK 2014). Die Freiwilligkeit ist im Kontext der externen stationären Qualitätssicherung nicht gegeben, dennoch kann die Überzeugung der Teilnehmer von Sinn und Nutzen des Verfahrens günstig beeinflusst werden.

Die *Überzeugung der Teilnehmer vom Nutzen des Verfahrens* gilt in der Literatur selbst als Erfolgsfaktor (BÄK 2014). Dieser steht in engem Zusammenhang mit den Erfolgsfaktoren *Kenntnis und Einhaltung der Verfahrensregeln durch alle Teilnehmer* (BÄK 2014). Sind die Teilnehmerinnen und Teilnehmer eines Verfahrens von diesem überzeugt, werden sie die Verfahrensregeln kennen und einhalten. Teilnehmerinnen und Teilnehmer des Verfahrens sind die Leistungserbringer, die Mitarbeiterinnen und Mitarbeitern der beauftragten Stelle sowie das Expertengremium. Existieren für die entsprechenden Akteure klare Verfahrensregeln und Voraussetzungen für die Qualitätsbewertung, führt das zu einer transparenteren und fairen Beurteilung der Versorgungsqualität und dadurch zu einer größeren Akzeptanz des Verfahrens. Zudem kann eine von der Qualitätsbewertung abgekoppelte Qualitätsförderung den Erfolg und die Akzeptanz der externen qualitätsfördernden Maßnahmen steigern. Insgesamt können im Rahmen der externen Qualitätssicherung dadurch möglichst optimale Bedingungen geschaffen werden, um das einrichtungsinterne Qualitätsmanagement beim Abbau von Qualitätsmängeln zu unterstützen.

3.8 Zusammenfassung

In den vorherigen Abschnitten wurden aus der wissenschaftlichen Literatur und den Methodischen Grundlagen V1.1 Kriterien für die Güte von Bewertungs- bzw. Beurteilungsprozessen und für den Erfolg von Qualitätsfördermaßnahmen hergeleitet. Ziel dieser Herleitung ist es, methodische Begründungen für die in den folgenden Kapiteln beschriebenen Empfehlungen für die Weiterentwicklung des Strukturierten Dialogs zu geben.

Aus dieser Herleitung ergibt sich, dass die quantitative Beurteilung der Versorgungsqualität mittels Indikatoren und die Beurteilung im Stellungnahmeverfahren unterschiedliche Rollen bei der Qualitätsbewertung haben, die aufeinander abgestimmt sein müssen. Quantitative Untersuchungsmethoden sind zu bevorzugen, wenn standardisierte Bewertungen eines Sachverhalts vorgenommen werden sollen (IQTIG 2019a, S. 42 f.), da sie eine hohe Objektivität ermöglichen und zumindest bei der Datenaus- und –bewertung ressourcenschonender sind als qualitative Untersuchungsmethoden. Dies ist der Fall in der externen Qualitätssicherung nach QSKH-RL, DeQS-RL und plan. QI-RL, wo einerseits Aussagen über die Erfüllung von Standards durch die Leistungserbringer im Mittelpunkt stehen, sowie die Ergebnisse für Auswahlentscheidungen von Patientinnen und Patienten Verwendung finden sollen.

Die qualitative Beurteilung im Stellungnahmeverfahren ist gemäß den hier hergeleiteten Kriterien nicht als ein eigenständiges Bewertungsinstrument zu verstehen, sondern sie erfüllt eine klar umschriebene Funktion im Gesamtverfahren (vgl. auch Abschnitt 2.1). Diese besteht in der Prüfung der Validität der Qualitätsaussage eines Qualitätsindikators in dem speziellen Fall eines Leistungserbringers, der auffällig wurde – in Abgrenzung zur Validität des Qualitätsindikators im Allgemeinen, die durch die Entwicklung entlang der Eignungskriterien sichergestellt wird. Dabei soll sich die Prüfung auf das Qualitätsmerkmal des Indikators und auf das durch den Referenzbereich operationalisierte Qualitätsniveau beziehen. Die Vorteile einer standardisierten Erhebung und Bewertung durch Qualitätsindikatoren gingen verloren bzw.

wären nicht erforderlich, wenn die Qualitätsindikatoren lediglich als Aufgreifkriterien zur Ermittlung auffälliger Leistungserbringer genutzt würden, das Stellungnahmeverfahren aber eine eigene, vom Qualitätsziel des Indikators unabhängige Bewertung vornähme.

Die aus dem Ziel der Beauftragung, ein möglichst einheitliches (objektives) Vorgehen im Stellungnahmeverfahren zu gewährleisten, abgeleiteten methodischen Anforderungen werden in den folgenden Kapiteln in Form von Empfehlungen für ein möglichst explizites Vorgehen ausgearbeitet. Wie in den vorangehenden Kapiteln deutlich wird, ergibt sich jedoch ein Zielkonflikt, wenn an ein qualitatives Verfahren, wie ein Stellungnahmeverfahren ähnlich hohe Anforderungen an die Objektivität angelegt werden wie an ein quantitatives Verfahren, wie die Indikatorbasierte Qualitätsmessung. Denn die zur Standardisierung des qualitativen Verfahrens notwendigen Maßnahmen erhöhen einerseits wenigstens teilweise den Aufwand für die Durchführung des Verfahrens und schränken andererseits notwendigerweise die Freiheit ein, unterschiedliche Informationen für das qualitative Verfahren heranzuziehen.

4 Eckpunkte des Rahmenkonzepts für die Qualitätsbewertung und -förderung

Ein zentrales Motiv für die Beauftragung zur Weiterentwicklung des Strukturierten Dialogs ist eine Steigerung der Objektivität und eine Reduktion des Aufwands (siehe Beauftragung S. 1). Wie in Abschnitt 3.3 dargelegt entspricht dieses Ziel im Verständnis des IQTIG einer stärkeren Standardisierung von Prozessen, die teilweise auch mit einer Aufwandsreduktion einhergeht. In der Analyse des bisher geführten Strukturierten Dialogs nach QSKH-RL wurden die drei Funktionen Qualitätsbewertung, Qualitätsförderung und Bewertung der Dokumentationsqualität identifiziert (siehe auch Abschnitt 2.1). Diese drei Funktionen sind auch in den Regelungen der DeQS-RL zu finden. Allerdings sind diese Bestandteile weder in der QSKH-RL noch in der DeQS-RL eindeutig voneinander abgegrenzt. Dialogische Elemente wie Begehungen und kollegiale Gespräche können beispielsweise gleichzeitig für die Bewertung der Versorgungsqualität und für die Vereinbarung von qualitätsverbessernde Maßnahmen eingesetzt werden. Durch die Vermischung beider Funktionen leiden jedoch die Objektivität und Effizienz von Qualitätsbewertung sowie die Effektivität der Qualitätsförderung. Beispielsweise ist es wahrscheinlich, dass Empfehlungen im Rahmen eines Peer Reviews weniger gut angenommen werden können, wenn gleichzeitig die Qualitätsbewertung noch aussteht. Demnach sollten Verbesserungspotenziale im kollegialen Gespräch erst identifiziert werden, wenn die Bewertung abgeschlossen ist (vgl. BÄK 2014: S. 35). Zudem könnte die Qualitätsbewertung dadurch beeinflusst werden, wie offen der Leistungserbringer für Vorschläge zur Qualitätsverbesserung ist. So könnte die Qualitätsbewertung beispielsweise strenger ausfallen, wenn sich ein Leistungserbringer Vorschlägen zur Qualitätsverbesserung verschließt, und weniger streng, wenn er sich offen für solche Vorschläge zeigt. All dies sind Beispiele für Einflüsse, die die Objektivität des Verfahrens reduzieren können (vgl. Abschnitt 3.3).

4.1 Modulare Betrachtungsweise

Aus diesen Gründen ist eine zentrale Eigenschaft im Rahmenkonzept für die beauftragte Weiterentwicklung die Trennung von Qualitätsbewertung und Qualitätsförderung: Es wird empfohlen, zukünftig den Prozess der Qualitätsbewertung sowohl konzeptuell als auch zeitlich vom Prozess der Qualitätsförderung abzugrenzen. Dies bedeutet, dass unter dem Begriff „Stellungnahmeverfahren“ ausschließlich die Prozesse von der Einholung bis zur abschließenden Bewertung der Stellungnahmen und des Indikatorergebnisses verstanden werden sollen. Auf die Qualitätsbewertung aufbauende Maßnahmen sollen nicht als Teil des Stellungnahmeverfahrens angesehen werden. Im Anschluss an die Bewertung bestehen neben der Qualitätsförderung prinzipiell verschiedene Handlungsoptionen (IQTIG 2019a: S. 38), z. B. im Rahmen von planungsrelevanten Qualitätsindikatoren oder Qualitätszu- und -abschlägen. Eine Trennung findet sich

ähnlich auch in die Regelungen zum Stellungnahmeverfahren nach § 11 der Richtlinie zu planungsrelevanten Qualitätsindikatoren (plan. QI-RL) ,[28] in der keine Fördermaßnahmen im Rahmen des Stellungnahmeverfahrens vorgesehen sind. Daher wird im vorliegenden Bericht die Unterscheidung der drei Funktionen des bisherigen Strukturierten Dialogs in Form von drei unterschiedlichen Modulen vorgeschlagen:

- Das **Modul „Qualitätsbewertung“** subsumiert alle Schritte, die von der Berechnung eines Indikatorergebnisses bis hin zu der abschließenden Qualitätsbewertung durchgeführt werden
- Das **Modul „Qualitätsförderung“** beinhaltet diejenigen Maßnahmen, die zur Steigerung der Versorgungsqualität bei einem Leistungserbringer nach abgeschlossener Qualitätsbewertung eingesetzt werden.
- Das **Modul „Daten- und Dokumentationsqualität“** beinhaltet das Vorgehen im Zusammenhang mit den sogenannten Auffälligkeitskriterien. Wie in Abschnitt 1.3 erläutert, ist dieses Modul nicht Teil der Beauftragung und daher auch nicht Teil des vorliegenden Konzepts.

Die folgenden Empfehlungen zielen darauf ab, dass im ersten Schritt aussagekräftige (valide) und nachvollziehbare (objektive) Qualitätsbewertungen vorgenommen werden, auf die weiterführende Maßnahmen, wie etwa qualitätssteigernde Maßnahmen bei den Leistungserbringern oder die Information von Patientinnen und Patienten über die Versorgungsqualität (Public Reporting), sinnvoll aufbauen können. Gemäß Beauftragung wird damit eine Standardisierung der Prozesse des Moduls „Qualitätsbewertung“ angestrebt. Die Empfehlungen orientieren sich an den methodischen Erfordernissen für aussagekräftige Qualitätsbewertungen (siehe Abschnitt 3.1).

4.2 Qualitätsindikatoren als kleinste Bewertungseinheit

Da ein Qualitätsindikator die kleinste Bewertungseinheit der gesetzlich verpflichtenden Qualitätssicherung ist, bezieht sich das Rahmenkonzept immer auf das Ergebnis eines einzigen Indikators. Dies bedeutet auch, dass sich die Qualitätsbewertung immer ausschließlich auf das Qualitätsmerkmal bezieht, das durch den betreffenden Indikator abgebildet werden soll. Eine Aussage über die „Gesamtqualität“ der Leistungserbringung einer Fachabteilung oder eines Standorts kann grundsätzlich nur in der Zusammenschau der Ergebnisse eines für diesen Zweck geeigneten QI-Sets getroffen werden (siehe auch Abschnitt 3.3).

Das grundsätzliche Vorgehen bei der Qualitätsbewertung gemäß des vorliegenden Konzepts ist identisch für alle Indikatortypen (Struktur, Prozess, Ergebnis) jeder Datenquelle (Dokumentation der Leistungserbringer, Sozialdaten, Daten aus Patientenbefragungen). Damit wird ein möglichst einheitliches Vorgehen für alle Indikatoren gewährleistet. Auch für so genannte Follow-up-Indikatoren, deren Auswertung sich auf Ereignisse (z. B. Revisionseingriffe oder der Tod einer Patientin oder eines Patienten) innerhalb eines bestimmten Beobachtungszeitraumes im Anschluss an ein interessierendes Ereignis (z. B. eine Organtransplantation) bezieht, kann und soll nach

[28] Richtlinie zu planungsrelevanten Qualitätsindikatoren gemäß § 136 Absatz 1 SGB V i. V. m. § 136c Absatz 1 und Absatz 2 SGB V. In der Fassung vom 15. Dezember 2016, zuletzt geändert am 18. Januar 2018, in Kraft getreten am 12. Mai 2018. URL: https://www.g-ba.de/informationen/richtlinien/91/ (abgerufen am 14.03.2019).

diesem Konzept verfahren werden. Denn Follow-up-Indikatoren unterscheiden sich nicht prinzipiell von anderen Indikatortypen. Fragen der Zuschreibbarkeit der Verantwortung und der Datenqualität, die sich bei Follow-up-Indikatoren besonders stellen, müssen im Rahmen der Eignungsbeurteilung der Indikatoren beantwortet werden (siehe auch Eignungskriterien, IQTIG 2019a, S. 150).

4.3 Grundlegender Ablauf des Moduls „Qualitätsbewertung“

In Abbildung 9 werden die Empfehlungen für die grundlegenden Prozesse der Module „Qualitätsbewertung“ und „Qualitätsförderung“ dargestellt und für das Modul „Qualitätsbewertung“ im Folgenden erläutert.

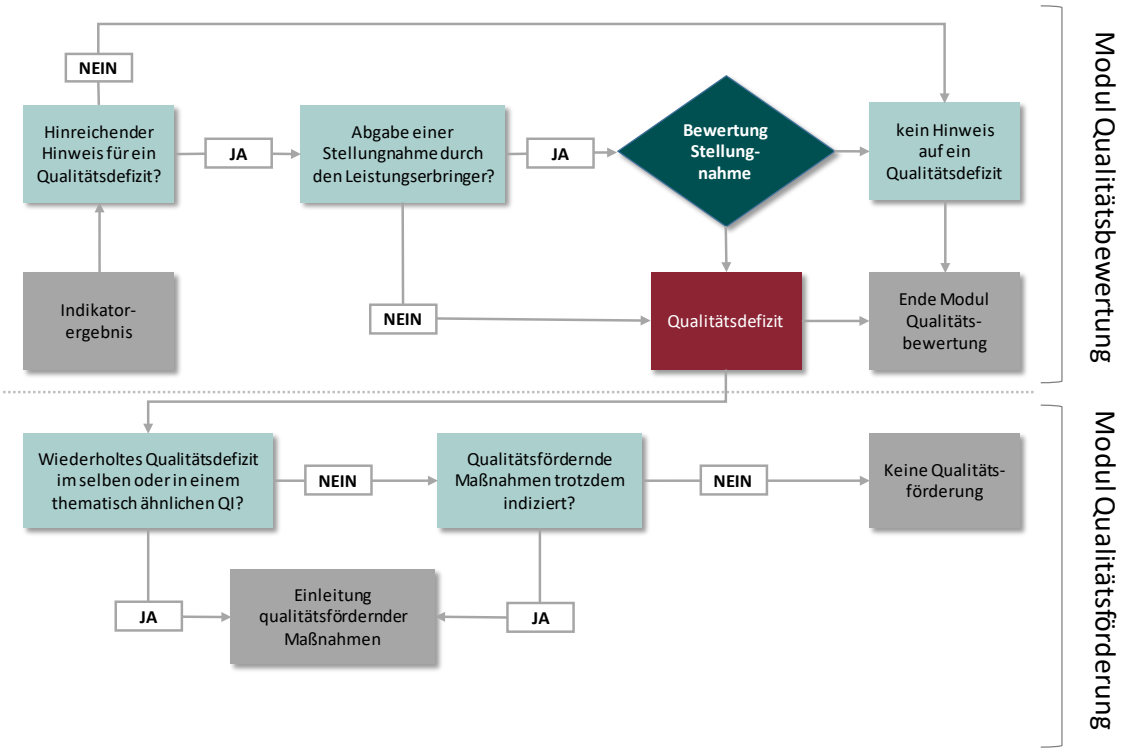

Abbildung 9: Grundlegender Ablauf der Qualitätsbewertung und -förderung

Legt ein Indikatorergebnis einen hinreichenden Hinweis für ein Qualitätsdefizit nahe, so ist dem Leistungserbringer von der LAG die befristete Möglichkeit zur schriftlichen Stellungnahme einzuräumen. Das zunächst abstrakte Kriterium des hinreichenden Hinweises für ein Qualitätsdefizit kann auf unterschiedliche Weise operationalisiert werden. Im Rahmen der QSKH-RL wurde bisher ein numerischer Vergleich zwischen Indikatorwert und Referenzbereichsgrenze vorgenommen („rechnerische Auffälligkeit“). Dieses Vorgehen wird jedoch für einige Prozess- und viele Ergebnisindikatoren als ungeeignet angesehen, da es mögliche stochastische, also fallzahlabhängige, Einflüsse auf das Indikatorergebnis unberücksichtigt lässt (vgl. Abschnitt 15.1, IQTIG 2019a). In Kapitel 2 wird daher das abstrakte Kriterium des hinreichenden Hinweises für ein Qualitätsdefizit unter Berücksichtigung stochastischer Einflüsse für die verschiedenen Indikatortypen (ratenbasierte Indikatoren, O/E-Indikatoren etc.) statistisch operationalisiert.

Mit der Aufforderung zur Stellungnahme wird dem Leistungserbringer die Möglichkeit eingeräumt, Einflussfaktoren auf das Indikatorergebnis, die nicht von ihm zu verantworten sind und nicht bereits im Qualitätsindikator oder dessen Risikoadjustierung oder mittels einer anderen statistischen Methode berücksichtigt sind, bei der Qualitätsbewertung geltend zu machen. Damit hat die Einholung und Bewertung einer Stellungnahme eine genau umschriebene Funktion, auf die hin die Kriterien für eine Einholung sowie für die fachliche Bewertung ausgerichtet sind. Im Rahmen der fachlichen Bewertung der Stellungnahme prüft eine Fachkommission die in der Stellungnahme angeführten Gründe für das Verfehlen des Referenzbereichs und beurteilt, ob diese tatsächlich nicht vom Leistungserbringer zu verantworten sind und damit der Hinweis aus dem Indikatorergebnis entkräftet ist. In Kapitel 6 werden der fachliche Bewertungsprozess sowie die Kriterien dafür genauer erläutert.

Die Abgabe einer Stellungnahme durch den Leistungserbringer soll freiwillig sein. Wird seitens des Leistungserbringers auf eine Stellungnahme verzichtet, gilt gemäß des Verständnisses von Qualitätsindikatoren (siehe Abschnitt 3.1), dass ein hinreichender Hinweis auf ein Qualitätsdefizit vorliegt, der Indikator in seiner Validität nicht angezweifelt wird und somit ein Qualitätsdefizit vorliegt (siehe Abbildung 9). Daher soll in diesem Fall die Bewertung „Qualitätsdefizit“ erfolgen.

Gemäß § 17 DeQS-RL kann die LAG auch bei „besonders guten Ergebnissen“ eine Stellungnahme anfordern. Hier wird jedoch empfohlen, die Einholung von Stellungnahmen auf Situationen zu beschränken, in denen ein hinreichender Hinweis auf ein Qualitätsdefizit gemäß den statistischen Kriterien besteht. Aus der Funktion von Stellungnahmen im Rahmen von *accountability* (vgl. Abschnitt 3.1) ergibt sich, dass Stellungnahmen für die Qualitätsbewertung nur dann indiziert sind, wenn die Validität des Indikatorergebnisses einer weiteren Prüfung unterzogen werden soll. Vor dem Hintergrund begrenzter Ressourcen bei den LAG und den Leistungserbringern wird daher empfohlen, diesen zweiten Prüfschritt nur im Sinne einer Steigerung der Spezifität des Vorgehens einzusetzen (s. u.).

Legt das Indikatorergebnis gemäß der empfohlenen biometrischen Auswertungsmethodik keinen hinreichenden Hinweis auf ein Qualitätsdefizit nahe, soll das Indikatorergebnis automatisch mit „kein Hinweis auf ein Qualitätsdefizit“ bewertet werden. Dies stellt eine begründete Asymmetrie im Vorgehen zugunsten der Leistungserbringer dar: Bei einem hinreichenden Hinweis auf ein Qualitätsdefizit auf Basis des quantitativen Indikatorergebnisses wird im Rahmen des Stellungnahmeverfahrens eine weitere Prüfung des Ergebnisses vorgenommen, während bei Abwesenheit eines solchen Hinweises kein weiterer Prüfschritt vorgesehen ist. Durch das Stellungnahmeverfahren wird somit die Spezifität des Vorgehens für „tatsächliche Qualitätsdefizite“ erhöht, im Vergleich zu einem Vorgehen ohne Stellungnahmeverfahren, da bei einem hinreichenden Hinweis für ein Qualitätsdefizit dieser Hinweis widerlegt werden kann, aber das Fehlen eines hinreichenden Hinweises für ein Qualitätsdefizit nicht widerlegt werden kann. Dies bedeutet, dass Einflüsse wie Fehler in der Datengrundlage und nicht vom Leistungserbringer zu vertretende Faktoren, die sich günstig für diesen auswirken (z. B. eine in diesem Jahr gesündere Patientenpopulation, die in der Risikoadjustierung nicht berücksichtigt ist), zu einem positiveren Indikatorergebnis führen können als die tatsächliche Versorgungsqualität hergibt.

Die Methodik der einzelnen Elemente und Prozessschritte wird in den folgenden Kapiteln ausführlich beschrieben.

4.4 Beschränkung der Informationsgrundlage auf Stellungnahmen

Die in der Beauftragung genannten Ziele einer möglichst einheitlichen Vorgehensweise – also einer hohen Objektivität – sowie einer Verschlankung des Verfahrens sind nur durch Standardisierung des Verfahrens erreichbar. Demgegenüber steht möglicherweise der Wunsch nach Berücksichtigung einer möglichst flexiblen Informationsgrundlage im zweiten, qualitativen Schritt. Durch (telefonische) Nachfragen, Begehungen, Präzisierungen der Stellungnahmen und die Möglichkeit korrigierte Stellungnahmen nachzureichen, kann die Informationsgrundlage nach Bedarf erweitert werden und damit möglicherweise die Wahrscheinlichkeit erhöht werden, bei inhaltlich zunächst unzureichenden Stellungnahmen doch noch entlastende Gründe zu finden (= Steigerung der Spezifität des Vorgehens, vgl. Abschnitt 2.1). Eine möglichst flexible Informationsgrundlage ist damit nur eingeschränkt vereinbar mit den in der Beauftragung genannten Zielen eines aufwandsarmen und objektiven Verfahrens. Erfolgt die Hinzuziehung und Bewertung zusätzlicher Informationen jedoch nicht nach einheitlichen Regeln, mindert dies die Objektivität des Verfahrens. Aus methodischer Sicht ist eine möglichst flexible Informationsgrundlage für das Stellungnahmeverfahren jedoch gar nicht notwendig, da wie in Abschnitt 3.1 hergleitet, die methodische Funktion des Stellungnahmeverfahrens darin gesehen wird, Einflussfaktoren zu identifizieren, die die Aussagekraft eines Indikatorergebnisses in Zweifel ziehen. Für diese Funktion erscheinen schriftliche Stellungnahmen, in denen Leistungserbringer aus ihrer Sicht vorliegende Einflussfaktoren benennen können, zweckmäßig und praktikabel. Des Weiteren erhöht sich durch eine flexiblere Informationsgrundlage auch der Aufwand und die Dauer des Verfahrens. Auch deshalb wird empfohlen, keine zusätzlichen Informationen über das Indikatorergebnis und die (kriterienkonforme) Stellungnahme hinaus für das Modul „Qualitätsbewertung“ hinzuzuziehen.

Im Gegensatz dazu sollte für das Modul „Qualitätsförderung“ eine möglichst breite und flexible Informationsgrundlage geschaffen werden können. Eine Notwendigkeit zur Standardisierung ergibt sich hier nicht, weder aus dem Auftrag noch aus methodischen Gründen, denn anders als die Ergebnisse von Qualitätsmessungen sind die Fördermaßnahmen sinnvollerweise nicht für Leistungserbringervergleiche vorgesehen. Im Gegenteil – wie in Kapitel 7 dargelegt – sollten Fördermaßnahmen auf die individuelle Situation des jeweiligen Leistungserbringers angepasst sein, um eine optimale Wirkung zu entfalten.

5 Statistische Auswertungsmethodik

Ziel dieses Kapitels ist die Darstellung eines einheitlichen Rahmenkonzepts zur statistischen Berechnung von Indikatorergebnissen anhand von Daten. Ein solches, übergreifendes Konzept für die externe Qualitätssicherung existiert bislang nicht, ist jedoch notwendig, um später fachlich zu verschiedenen Methoden der rechnerischen Auffälligkeitseinstufung diskutieren zu können. Diese umfassende Darstellung der Auswertungsmethodik bietet den notwendigen Rahmen, um fundierte Vorschläge für die Operationalisierung des Begriffs „hinreichender Hinweis auf ein Qualitätsdefizit“ aus Kapitel 4 zu generieren.

Betrachtet wird in diesem Kapitel ein bestehender Qualitätsindikator, welcher entsprechend den Schritten in den „Methodischen Grundlagen“ des IQTIG (IQTIG 2019a) entwickelt worden ist und der nun im Rahmen der DeQS-RL berechnet werden soll. Die statistische Auswertungsmethodik für die Berechnung von QI-Ergebnissen wird im Folgenden grundlegend dargestellt: dies enthält den prinzipiellen Aspekt der Berechnung von QI-Ergebnissen anhand der beobachteten Daten und den spezielleren Aspekt der Auslösung des Stellungnahmeverfahrens für Leistungserbringer anhand deren QI-Ergebnisse im Rahmen einer statistischen Klassifikation. Fokus des vorliegenden Kapitels ist dieser Klassifikationsaspekt.

Der Abschnitt 5.1 liefert ein Rahmenkonzept für die statistische Auswertungsmethodik für Qualitätsindikatoren. Danach richtet sich der Fokus in Abschnitt 5.2 auf den Klassifikationsaspekt der Auswertungsmethodik, welcher formal im Rahmen eines Entscheidungsproblems dargestellt wird. Der Abschnitt 5.3 beschreibt dann verschiedene Formen der rechnerischen Auffälligkeitseinstufung für die Daten eines Erfassungsjahres. Dabei werden die Methoden statistisch formal hergeleitet und jeweils das Entscheidungssetting diskutiert, in dem die Einstufungsmethode die optimale Klassifikationsgüte hat. Abschnitt 5.4 erweitert diese Methoden dann, damit Daten aus mehreren Erfassungsjahren bei der Auffälligkeitseinstufung berücksichtigt werden können. Insgesamt sind die Abschnitte 5.2 bis 5.4 anspruchsvoller statistischer Natur und setzen fortgeschrittene Kenntnisse inferenzstatistischer Methoden voraus. Eine Darstellung auf diesem Niveau ist jedoch notwendig, um später fachlich fundiert über die Optimierung der Effizienz des SD-Prozesses diskutieren zu können. Eine Illustration der Einstufungsmethodik findet sich im Abschnitt 5.5 und eine etwas allgemeinverständlichere Zusammenfassung der Inhalte der Abschnitte 5.2 bis 5.4 im Abschnitt 5.6.

5.1 Rahmenkonzept der Auswertungsmethodik

Jede neu entwickelte statistische Auswertungsmethodik im Bereich der externen Qualitätssicherung sollte eine belastbare quantitative Aussage über die Qualität der Leistungserbringer liefern. Die Ergebnisse der statistischen Auswertung sollen hilfreiche Informationen für Patientinnen und Patienten und für die Leistungserbringer zur Verfügung stellen und damit eine erhöhte Transparenz über die Qualität der Versorgung schaffen. Damit die Methodik sinnvoll zum Einsatz kommen kann, muss sichergestellt werden, dass sie für die vorliegenden bzw. erwarteten Fallzahlen, in Kombination mit den interessierenden Fragestellungen, geeignet ist.

Grundlage für die Auswertung sind die Daten, die gemäß der Spezifikation für das zugehörige QS-Verfahren im Rahmen der DeQS-RL erhoben werden. In der einfachsten Form wird der Indikator durch Mengendefinitionen der zu betrachtenden Population und des interessierenden Ereignisses in diesen Daten operationalisiert. Ein wichtiger Bestandteil dieser Mengendefinitionen ist die Festlegung, aus welchen Entitäten[29] die Population besteht, dies könnten z. B. Patientinnen und Patienten, aber auch Prozeduren sein. Prinzipiell könnte die Entität auch ein Fragebogen aus einer Einrichtungsbefragung sein. Im Folgenden wird jedoch der Begriff „Fall" als generischer Begriff für die betrachtete Entität verwendet.

Grundsätzlich sind folgende weitere Festlegungen für eine konkrete Auswertung eines Indikators zu treffen: Die betrachtete Gruppierungseinheit für die Auswertung sowie der zu bewertende Zeitraum. Prinzipiell wird die Auswertung für mehrere Gruppierungseinheiten vorgenommen, z. B. auf Ebene der Leistungserbringer (entlassender oder leistungserbringender Standort), der Landesebene und der Bundesebene. Für die Auslösung des Stellungnahmeverfahrens ist jedoch nur die Einheit Leistungserbringer relevant, weshalb andere Auswertungseinheiten im Folgenden nicht weiter diskutiert werden. Der zu bewertende Zeitraum ist der temporale Aspekt der Datengrundlage. Eine klare Definition der Datengrundlage für die Auswertung wird dabei vorausgesetzt, es geht hier nur um den zeitlichen Aspekt der Datengrundlage. In den einfachsten Fällen ist dies das Erfassungsjahr, d. h. z. B. alle Entlassungen in der betrachteten Gruppierungseinheit im Kalenderjahr X. Im Rahmen der gleichen Auswertung kann es durchaus zu mehreren unterschiedlichen Berechnungen des gleichen Indikators bzgl. unterschiedlicher Gruppierungseinheiten und Zeiträume kommen, z. B. Leistungserbringerergebnisse im Vergleich zum Bundesergebnis, Ergebnis mit den aktuellen Rechenregeln für das aktuelle Erfassungsjahr vs. Ergebnis mit aktuellen Rechenregeln für das vorherige Erfassungsjahr.

Die quantitative Auffälligkeitseinstufung[30] der Leistungserbringer, für die ein Stellungnahmeverfahren eingeleitet werden soll, sollte nachvollziehbar sein und erlauben, dass daraus Schlüsse gezogen und Konsequenzen abgeleitet werden können. Grundlegende Voraussetzung dafür und für die Akzeptanz der Auswertungsergebnisse ist, dass die Methodik fachlich fundiert ist und zugrunde liegende Annahmen explizit gemacht werden, damit die Adäquatheit der Annahmen und deren Konsequenzen diskutiert werden können. Des Weiteren ist die Methodik so konkret darzustellen, dass die Leistungserbringer die entsprechenden Ergebnisse nachvollziehen können. Die quantitative Einstufung soll zudem eine hohe Treffsicherheit[31] gewährleisten, um einen

[29] Entität wird hier als Sammelbegriff für die Eigenschaften der Fälle verwendet, die Gegenstand der Qualitätssicherung sind. Die Festlegung der Entität ist vor allem für QS-Verfahren mit Unterbogenstruktur relevant, da hier die Festlegung der Entität die Zählweise bei der Bestimmung der Populationen aus den zusammengefügten Bögen festlegt. Als Beispiel könnte Entität in der Geburtshilfe entweder die Mütter oder die Kinder bezeichnen.

[30] In diesem Kapitel wird der Begriff „quantitative Auffälligkeit" als Überbegriff für verschiedene statistische Verfahren zur Klassifikation des Leistungserbringerergebnisses verwendet. Es wird bewusst ein neuer Begriff hierfür verwendet, um sich vom Begriff der QSKH-RL, der „rechnerischen Auffälligkeit", zu unterscheiden, welches den direkten Vergleich zwischen Indikatorergebnis und Referenzbereich vorschreibt, d. h. bei QIs mit unerwünschten interessierenden Ereignissen, ob Indikatorergebnis > Referenzwert.

[31] Mit „Treffsicherheit" ist die Genauigkeit der Klassifikation gemeint, d. h. die Eigenschaft des Klassifikationsverfahrens, tatsächliche Qualitätsdefizite anhand der QI-Ergebnisse zu identifizieren und umgekehrt auch das Fehlen tatsächlicher Qualitätsdefizite als solche zu identifizieren.

effizienten Verfahrensablauf zu sichern, bei dem jedoch keine Qualitätsdefizite übersehen werden. Die oben erwähnten Kriterien machen schnell klar, dass die gewählte Methodik sich im Spannungsfeld zwischen anspruchsvollen Berechnungen und Kommunizierbarkeit befinden, zwischen denen behutsam abgewogen werden muss.

Für die Wahl einer geeigneten statistischen Auswertungsmethodik spielen sowohl die indikatorübergreifende Herangehensweise und Stichprobenart sowie die Berechnungsart des betrachteten Indikators eine wichtige Rolle. Zusammen mit der Bewertungsart stellen die Festlegungen eine Taxonomie für die Auswertungsmethodik eines Qualitätsindikators dar und sind vorab für jede Auswertung des Qualitätsindikators festzulegen. Im Folgenden werden Herangehensweise, Stichprobenart für die vorgegebene Auswertungsebene und Berechnungsart kurz vorgestellt, da sie die Grundlage für eine fundierte Diskussion der Bewertungsart als Klassifikation der Leistungserbringerergebnisse sind.

5.1.1 Herangehensweise

Hinsichtlich des Zwecks einer statistischen Datenauswertung lassen sich zwei Herangehensweisen – die analytische und die enumerative – differenzieren (vgl. z. B. Deming (1953)). Die folgende Erläuterung dieser beiden Herangehensweisen wurde (in modifizierter Form) bereits im Rahmen der Abschlussberichte zur Entwicklung von Patientenbefragungen der Qualitätssicherungsverfahren *Perkutane Koronarintervention (PCI) und Koronarangiographie* sowie *Schizophrenie* angeführt (IQTIG 2018c, IQTIG 2018d). Die analytische Herangehensweise wurde auch implizit bei der Entwicklung der statistischen Auswertungsmethodik für die plan. QI-RL vorausgesetzt (IQTIG 2016).

Analytische Herangehensweise

Bei der analytischen Herangehensweise liegt das Interesse am zugrunde liegenden Prozess. Im Rahmen der externen Qualitätssicherung bedeutet dies, dass im Gegensatz zum reinen Auszählen der dokumentierten interessierenden Ereignisse in einem Erfassungsjahr anhand eines statistischen Modells Schlüsse über die Systematik bezüglich der betrachteten Qualitätsmerkmale beim Leistungserbringer gezogen werden sollen. Das Interesse liegt somit in der *zugrunde liegenden Kompetenz* des Leistungserbringers bezüglich des im Indikator abgebildeten Qualitätsmerkmals, welches durch die Rechenvorschrift des Indikators statistisch operationalisiert wird. Dabei muss es sich nicht um eine spezifische, benennbare Kompetenz handeln. Vielmehr abstrahiert der im statistischen Modell angenommene zugrunde liegende Kompetenzparameter in diesem Zusammenhang das komplexe Zusammenspiel vieler Eigenschaften des Leistungserbringers, die sich zusammen mit weiteren Faktoren, wie beispielsweise patientenseitigen Faktoren, auf die in dem Indikator betrachtete Behandlung der Patientinnen und Patienten auswirken. Bei einem Ratenindikator stellt der Kompetenzparameter die zugrunde liegende Wahrscheinlichkeit, mit der das interessierende Ereignis bei den Fällen der betrachteten Grundgesamtheit eintritt, dar. Patientenseitige Faktoren bleiben bei Ratenindikatoren unberücksichtigt. Da die Operationalisierung immer ein Kompromiss zwischen Präzision und Aufwand der Datenerhebung darstellt und darüber hinaus i. d. R. nicht alle für den Indikator relevanten Informationen erhoben werden können, ist der Kompetenzparameter immer im Kontext der Operationalisierung,

d. h. der formulierten Rechenregeln des Indikators zu interpretieren. Zum Beispiel können besondere Versorgungssituationen oft nicht in den Rechenregeln des Indikators abgebildet werden. Somit kann der Kompetenzparameter des verwendeten statistischen Modells diesen Aspekt der „Kompetenz“ auch nicht berücksichtigen. Diese Berücksichtigung ist stattdessen Aufgabe im Zuge der fachlichen Bewertung. Für die Berechnung und Bewertung der Ergebnisse wird anhand von statistischer Inferenz basierend auf den zu diesem Zweck erhobenen Daten eines Erfassungsjahres mittels der Rechenvorschriften der Indikatoren jeweils ein Schätzwert für den jeweiligen latenten Kompetenzparameter des Leistungserbringers ermittelt. Diese Schätzwerte, d. h. die tatsächlich gemessenen Ergebnisse der Leistungserbringer in den Qualitätsindikatoren können als komplexe Funktion verschiedener Einflussgrößen aufgefasst werden (IQTIG 2019a: 181). Es wird angenommen, dass es unbekannte oder nicht erfasste Einflüsse auf die tatsächlich gemessenen Ergebnisse der Leistungserbringer gibt, die im Folgenden unter dem Konzept der Stochastizität subsumiert werden (vgl. auch IQTIG (2019a)). Da zur Bewertung der Leistungserbringer nur durch ihn beeinflussbare Einflüsse eingehen sollen (IQTIG 2017d, Kapitel 10), wird die statistische Unsicherheit resultierend aus der Stochastizität bei der Bewertung berücksichtigt. Bei der analytischen Herangehensweise wird, im Gegensatz zur enumerativen Herangehensweise, somit selbst dann statistische Unsicherheit angegeben und bei der Bewertung von Leistungserbringerergebnissen berücksichtigt, wenn für alle Fälle eines Leistungserbringers im betrachteten Erfassungsjahr Dokumentationen vorliegen. Da bei der analytischen Herangehensweise keine Schlüsse über eine endliche Population gezogen werden, sondern eine Aussage über den zugrunde liegenden Prozess getroffen wird, ist weniger entscheidend, ob es sich um eine Stichprobe oder eine Vollerhebung im klassischen Sinn handelt. Vielmehr ist wichtig, dass eine ausreichende Anzahl an Fällen vorliegt um eine präzise Schätzung des zugrunde liegenden Kompetenzparameters – also der Systematik – zu ermöglichen. Die zugrunde liegenden Kompetenzparameter der Indikatoren werden jeweils anhand der für ein Erfassungsjahr vorliegenden Daten und basierend auf mathematisch-statistischen Annahmen an den zugrunde liegenden stochastischen Prozess geschätzt, wirken sich letztlich jedoch auch auf die Fälle über das Ende des Erfassungsjahres hinaus aus, sofern angenommen werden kann, dass sich die zugrunde liegende Kompetenz für den jeweiligen Indikator nicht oder nur langsam verändert. Für die analytische Herangehensweise gibt die Einteilung in Erfassungsjahre lediglich vor, welche Daten zur Schätzung der Kompetenzparameter verwendet werden. Es ist jedoch nicht davon auszugehen, dass sich die zugrunde liegende Kompetenz abrupt nach Abschluss des Erfassungsjahres verändert, sodass sich annehmen lässt, dass sie sich auch auf zukünftig behandelte Patientinnen und Patienten auswirkt.

Die analytische Herangehensweise ist insbesondere dann angebracht, wenn Verallgemeinerungen über das Geschehene hinaus von Interesse sind (Deming 1953). Diese Herangehensweise findet in einigen internationalen Arbeiten im Rahmen der externer Qualitätssicherung Anwendung (siehe z. B. Spiegelhalter et al. 2012, Ash et al. 2012)[32] und bildet die Basis für die Auswertungen von Qualitätsindikatoren im Rahmen der QSKH-RL, der DeQS-RL und der plan. QI-RL, da

[32] Insbesondere kann immer von einer analytische Herangehensweise ausgegangen werden, wenn trotz Vollerhebung im klassischen Sinn statistische Unsicherheit berücksichtigt wird.

es für die meisten Prozessindikatoren und alle Outcome-Indikatoren nicht um das reine Zahlenwesen im Sinne eines statistischen Jahrbuchs geht, sondern darum, Rückschlüsse auf die Qualitätsprozesse des Leistungserbringers zu ziehen, um die Ergebnisse als Teil der Qualitätssicherung und zur Patienteninformation prospektiv einzusetzen. Die Berücksichtigung stochastischer Einflüsse bei analytischer Herangehensweise lässt sich umsetzen, indem die daraus resultierende statistische Unsicherheit nicht ignoriert, sondern in der Bewertung der Kompetenz explizit berücksichtigt wird. Eine Möglichkeit, statistische Unsicherheit zu berücksichtigen, ist, neben dem Schätzwert für den zugrunde liegenden Kompetenzparameter auch ein zugehöriges Unsicherheitsintervall anzugeben und dieses bzw. einen dazugehörigen statistischen Test auch in der Einstufung der Leistungserbringer zu verwenden (IQTIG 2019a). Dies führt zu einer Unterteilung der statistischen Auswertungsmethodik in drei Bestandteile:

1. Indikatorwert
2. zugehöriges Unsicherheitsintervall
3. quantitative Auffälligkeitseinstufung

Im Rahmen der analytischen Herangehensweise stellt der Indikatorwert im Allgemeinen einen Punktschätzer eines im Rahmen eines statistischen Modells definierten, zugrunde liegenden Parameters des Leistungserbringers dar (IQTIG 2019b, Kapitel 12). Dieser Parameter ist zunächst unbekannt und wird aus den (für diesen Indikator relevanten) vorliegenden Daten für den Leistungserbringer geschätzt. Vereinfachend wird dieser latente Parameter im Folgenden auch der QI-spezifische Kompetenzparameter des Leistungserbringers genannt. Da verschiedene Indikatoren ggf. eine Aussage über die Kompetenz der Leistungserbringer bezüglich unterschiedlicher Patientenpopulationen bzw. Behandlungsarten ermöglichen sollen, werden für jeden Indikator Bedingungen formuliert, unter denen die Patientinnen und Patienten in die Berechnung eingeschlossen werden. Patientinnen und Patienten, die diese Bedingungen erfüllen, werden im Folgenden auch als Teil der Grundgesamtheit, d. h. der Zielpopulation des Indikators, bezeichnet, wobei diese klar von dem Konzept der endlichen Grundgesamtheit in der enumerativen Herangehensweise zu unterscheiden ist. Diese Tatsache zeigt erneut, dass der Kompetenzparameter immer im Kontext der formulierten Rechenregeln des Indikators zu interpretieren ist.

Eine Berücksichtigung der oben beschriebenen stochastischen Einflüsse, die im Rahmen der analytischen Herangehensweise angenommen werden, findet im Rahmen von parametrischen statistischen Modellen über Verteilungsannahmen für die beobachtete Anzahl an interessierenden Ereignissen statt. Konkret bedeutet dies, dass die Anzahl an interessierenden Ereignissen als Zufallsvariable modelliert wird, dessen Verteilung von einem zugrunde liegenden Parameter abhängt, der als zugrunde liegende Kompetenz des Leistungserbringers im betrachteten QI interpretiert werden kann. Die Berechnung des Indikatorwerts für einen Leistungserbringer kann somit als eine Schätzung – im inferenzstatistischen Sinne – für die zugrunde liegende Kompetenz des Leistungserbringers im betrachteten QI gesehen werden. Da dieser statistische Schätzwert für die Kompetenz nicht die Unsicherheit bei der Bestimmung enthält, sollte zu jedem Indikatorwert auch ein zugehöriges Unsicherheitsintervall bestimmt werden, dessen Breite angibt, wie belastbar der berechnete Indikatorwert im Rahmen der Annahmen ist.

Allgemein dienen Unsicherheitsintervalle der Quantifizierung von statistischer Unsicherheit bezüglich der aus den Daten gewonnenen Informationen über den unbekannten Parameter von Interesse. Unter den getroffenen Annahmen über den datengenerierenden Prozess gibt die Breite eines Unsicherheitsintervalls an, wie stark die stochastischen Einflüsse bei der Bestimmung des Schätzwertes ins Gewicht fallen. Je schmaler das Intervall, desto belastbarer sind die Informationen, die aus den erhobenen Daten über den zugrunde liegenden Parameter gewonnen werden können. Die Breite von Unsicherheitsintervallen hängt u. a. von der Fallzahl ab, die dem Schätzwert zugrunde liegt. Die Verwendung von Unsicherheitsintervallen beschränkt sich nicht nur auf die reine Information über statistische Unsicherheit: die dahinterliegende Unsicherheit wird in der quantitativen Auffälligkeitseinstufung explizit berücksichtigt – teilweise direkt über die Unsicherheitsintervalle.

Bei der quantitativen Auffälligkeitseinstufung der Leistungserbringer handelt es sich um eine statistische Klassifikation. Ziel ist es, die Qualität der Leistungserbringer anhand der vorliegenden Daten auf Auffälligkeiten zu prüfen und im Falle von Auffälligkeiten Stellungnahmeverfahren einzuleiten, die dann ggf. Maßnahmen zur Qualitätsverbesserung initiieren. Da sich die der quantitativen Auffälligkeit anschließenden Prozesse sowohl in deren Konsequenzen als auch in Art und Umfang in einem sensiblen Kontext befinden, ist es notwendig, dass die quantitative Klassifikation treffsicher ist. Zur Definition und Bestimmung der Treffsicherheit, d. h. der Sensitivität und Spezifität einer statistischen Klassifikation, wird jedoch ein Außenkriterium, ein sogenannter Goldstandard, benötigt, den es in diesem Kontext nicht wirklich gibt (vgl. Seite 49 in IQTIG (2019b)).

Basiert die Einstufung ausschließlich auf dem berechneten Indikatorwert eines Leistungserbringers, so besteht die Gefahr, dass insbesondere wenn wenig Information über die interessierende Größe vorhanden ist, zufällige Einflüsse und Konstellationen dazu führen, dass eine Fehlentscheidung getroffen wird. Eine Fehleinschätzung kann dabei in beide Richtungen, d. h. falsch positiv bzw. falsch negativ, erfolgen. Um dies zu vermeiden und zu einer belastbaren Aussage zu gelangen, sollte die statistische Unsicherheit in der Bestimmung des Indikatorwerts bei der quantitativen Einstufung berücksichtigt werden.

Enumerative Herangehensweise

Im Gegensatz dazu liegt das Interesse bei der *enumerativen* Herangehensweise in einer interessierenden Größe in einer endlichen Grundgesamtheit, beispielsweise gegeben durch die vom Leistungserbringer im betrachteten Erfassungsjahr behandelten Patientinnen und Patienten, die die Grundgesamtheitsbedingungen erfüllen. Es wird angenommen, dass die interessierende Größe in der Grundgesamtheit deterministisch ist, d. h. es keine unbekannten oder nicht erfassten Prozesse gibt, die die Messung beeinflussen. Im Kontext der enumerativen Herangehensweise ist es daher im Gegensatz zur analytischen Herangehensweise sinnvoll zu unterscheiden, ob es sich bei der Stichprobenart (siehe nächster Abschnitt) um eine Vollerhebung handelt oder nicht. Handelt es sich um eine Vollerhebung, so wird keine statistische Unsicherheit berücksichtigt und das beobachtete Indikatorergebnis des Leistungserbringers wird direkt mit dem Referenzwert verglichen. Liegt das Leistungserbringerergebnis über dem Referenzwert, dann ist der Leistungserbringer quantitativ auffällig. Liegt hingegen keine Vollerhebung vor, so ist das Ziel,

Rückschlüsse von der Stichprobe auf die endliche Grundgesamtheit zu ziehen, und es muss die statistische Unsicherheit, die durch das Betrachten der Stichprobe entsteht, berücksichtigt werden, um korrekte Inferenz zu betreiben (Kauermann und Küchenhoff 2011). Die enumerative Herangehensweise ist dann angebracht, wenn das Ziel der Auswertung ist, eine Bestandsaufnahme zu machen und die interessierende, deterministische Größe für eine vorgegebene, endliche Population zu beschreiben. Diese Herangehensweise liegt beispielsweise den Auswertungen zu besonders häufigen Dokumentationsfehlern im Rahmen der Datenvalidierung nach QSKH-RL zugrunde (vgl. Kapitel 2 im Abschlussbericht zu „Kriterien für den gezielten Datenabgleich in der Datenvalidierung nach QSKH-RL“ (IQTIG 2018f)).

Zusammenfassend liegen die statistisch-methodischen Herausforderungen in der Auswertungsmethodik hauptsächlich in der analytischen Herangehensweise bzw. in der enumerativen Herangehensweise, wenn (z. B. aus Effizienzgründen) die vorliegenden Daten nur Stichproben der eigentlich interessierenden endlichen Grundgesamtheit ausmachen.

5.1.2 Stichprobenart für die vorgegebene Auswertungsebene

Im Rahmen der externen Qualitätssicherung finden bereits unterschiedliche Stichprobenarten Anwendung. Für die meisten Verfahren der stationären Qualitätssicherung liegen Vollerhebungen vor, weil alle Leistungserbringer, die im betrachteten Erfassungszeitraum Patientinnen und Patienten behandeln, die durch den QS-Filter ausgelöst werden, QS-Dokumentationsdaten zu all diesen Fällen liefern müssen. Hingegen werden im Rahmen der Datenvalidierung und zukünftig auch im Rahmen der Patientenbefragungen Zufallsstichproben von Leistungserbringern bzw. Patientinnen und Patienten beim Leistungserbringer gezogen.

Auch wenn die Unterscheidung, ob eine Vollerhebung vorliegt oder nicht, bei der analytischen Herangehensweise keine Rolle spielt, ist die Art der Stichprobenziehung dennoch für die statistische Inferenz relevant. Wichtig bei der Betrachtung der Stichprobenart ist, dass sie von der Auswertungsebene abhängt. Im Fall der Patientenbefragungen beispielsweise handelt es sich pro Leistungserbringer um eine einfache Zufallsstichprobe. Ist jedoch eine Auswertung aller Fragebögen auf Bundesebene von Interesse, so muss beachtet werden, dass die Gesamtheit der Fragebögen nicht als zufällige Stichprobe aus allen behandelten Patientinnen oder Patienten interpretiert werden kann, weil unterschiedlich viele Fragebögen pro Leistungserbringer eingehen und diese ggf. auch nicht den gleichen Anteil an jeweils bei den Leistungserbringern behandelten Patientinnen und Patienten darstellen.

5.1.3 Berechnungsart

Die Berechnungsart eines Indikators gibt an, wie das Indikatorergebnis (Punktschätzer) konkret basierend auf den Mengendefinitionen der Grundgesamtheit und der interessierenden Ereignisse sowie ggf. zusätzlicher Informationen (wie Patienteneigenschaften bei risikoadjustierten Indikatoren) bestimmt wird. Beispielsweise gibt die Berechnungsart „Rate“ an, dass der zugrunde liegende Kompetenzparameter als Anteil an Fällen mit interessierendem Ereignis an Fällen in der Grundgesamtheit berechnet wird. Neben Indikatoren mit Bewertungsart „Rate“, die in den bisherigen Auswertungen im Rahmen der externen Qualitätssicherung den größten Anteil

ausmachen, finden auch weitere Berechnungsarten Verwendung, wie beispielsweise das standardisierte Mortalitätsverhältnis (SMR, auch oft einfach nur O/E genannt), Kaplan-Meier-adjustierte Rate, standardisiertes Inzidenzverhältnis (SIR) und verschiedene Index-Konstruktionen.

5.1.4 Bewertungsart

Für die Bewertung von Indikatorergebnissen eines Leistungserbringers ist zusätzlich die Bewertungsart entscheidend, welche die statistische Klassifikationsmethode für den Vergleich mit dem Referenzbereich unter Berücksichtigung von Herangehensweise, Stichprobenart und Berechnungsart festlegt und als Operationalisierung des Begriffs „hinreichender Hinweis auf ein Qualitätsdefizit“ verstanden werden soll. Um hervorzuheben, dass es sich bei dem betrachteten Hinweis um statistische Evidenz aus den Ergebnissen des Qualitätsindikators anhand der vorliegenden Daten handelt, wird für die Klassifikation in diesem Bericht, wie oben erläutert, der Überbegriff „quantitative Auffälligkeitseinstufung“ verwendet.

5.1.5 Ziele der Auswertung für einen Leistungserbringer

Zusammenfassend stellen die vier vorgestellten Komponenten Herangehensweise, Stichprobenart für die vorgegebene Auswertungsebene, Berechnungsart und Bewertungsart eine Taxonomie für die Auswertungsmethodik eines Qualitätsindikators dar, und sind vorab für jede Auswertung des Qualitätsindikators festzulegen. Im Folgenden wird nun im Rahmen dieser Taxonomie auf die relevanten Ziele für die Berechnung und Bewertung für Leistungserbringer bei den Auswertungen fokussiert. Folgende Ziele sollen mit der Auswertung für einen Leistungserbringer mit den vorliegenden Daten erreicht werden:

- Zahlenergebnisse, die im Rahmen eines Public Reporting Transparenz über das Versorgungsgeschehen des Leistungserbringers herstellen können (vgl. Kapitel 11 in IQTIG (2019a)) oder für andere Verwendungszwecke weiterverwendet werden können. Dies wird im Folgenden das Leistungserbringerergebnis des QI genannt. Für einen Ratenindikator besteht das Leistungserbringerergebnis z. B. aus: Anzahl an Fällen in der Grundgesamtheit des Leistungserbringers im QI für das relevante Erfassungsjahr, Anzahl an interessierenden Fällen, QI-Wert und das zugehörige Unsicherheitsintervall.
- Eine zusätzliche quantitative Bewertung des beobachteten QI-Ergebnisses des Leistungserbringers für den zu bewertenden Zeitraum als Klassifikation in einer der beiden Kategorien „quantitativ unauffällig“ bzw. „quantitativ auffällig“. Dabei bedeutet „quantitativ auffällig“, dass das QI-Ergebnis hinreichende statistische Hinweise auf ein Qualitätsdefizit enthält, sodass diese im Sinne des Moduls „Qualitätsbewertung“ in Kapitel 4 im Rahmen eines Stellungnahmeverfahrens zu prüfen ist. Analog bedeutet „quantitativ unauffällig“, dass nicht genug bzw. keine Hinweise für ein solches Qualitätsdefizit vorliegen. Die quantitative Einstufung ist somit als erster Schritt eines Prozesses zu sehen, welcher im zweiten Schritt quantitativ auffälligen Leistungserbringern die Möglichkeit zur Reflexion ihrer Kompetenz und ggf. zur Erklärung dieser quantitativen Auffälligkeit im Rahmen einer Stellungnahme bietet, die dann fachlich bewertet wird.

Obwohl die Zahlen des Public Reportings nicht zwingend die gleiche Datengrundlage verwenden müssen[33] wie die quantitative Auffälligkeitseinstufung, ist es wichtig, dass die beiden Ziele mit einem einheitlichen statistisch-methodischen Ansatz verfolgt werden. Dies ist notwendig, um spätere Inkonsistenzen soweit möglich zu vermeiden, z. B. wenn im Rahmen eines Stellungnahmeverfahrens über die Ergebnisse der einzelnen Jahre diskutiert wird. In Abschnitt 5.1 wird daher ein einheitlicher Rahmen dargestellt, welcher die Komponenten der statistischen Auswertung eines Qualitätsindikators benennt und kategorisiert. Da für die vorliegende Beauftragung aus statistischer Sicht vor allem die Methodik zur quantitativen Auffälligkeitseinstufung von Interesse ist, wird diese in Abschnitt 5.2 in den Gesamtrahmen des Bewertungsprozesses eingebettet. Es werden verschiedene Vorschläge für eine Methodik für die quantitative Auffälligkeitseinstufung gemacht.

Das Ziel der nächsten Abschnitte ist die Entwicklung einer mathematisch-statistischen Methodik, welche fair und transparent und ggf. unter Berücksichtigung stochastischer Unsicherheit die Ergebnisse eines Leistungserbringers mit dem Referenzwert vergleicht. Dies soll die Basis darstellen für die Entscheidung, ob anhand der beobachteten Daten hinreichende statistische Hinweise auf ein Qualitätsdefizit vorliegen.

Bei der Operationalisierung einer solchen Methodik ist zu bedenken, dass jede quantitative Auffälligkeit prinzipiell einen qualitativen Prozess nach sich ziehen wird. Der Vorschlag aus Kapitel 4 ist es, diesen qualitativen Prozess im Rahmen eines Stellungnahmeverfahrens durchzuführen, vgl. Abbildung 9. Zusammenfassend kann der Bewertungsprozess daher als zweistufiges Verfahren mit den folgenden beiden Stufen angesehen werden:

- Schritt 1: Es werden Leistungserbringer identifiziert, bei denen anhand der vorliegenden Daten genügend statistische Hinweise auf ein Qualitätsdefizit bestehen (quantitative Auffälligkeitseinstufung).
- Schritt 2: Nur wenn die beobachteten Ergebnisse zu einer solchen statistisch belastbaren Abweichung vom Referenzwert des betrachteten Qualitätsindikators führen, kommt es zur Initiierung eines Stellungnahmeverfahrens wie es in Kapitel 4 beschrieben wird, welches die statistischen Hinweise überprüft und entsprechend fachlich bewertet (qualitativer Schritt).

Damit diese qualitative Überprüfung sorgfältig, einheitlich, transparent und fair durchgeführt werden kann, sind entsprechende Ressourcen für Schritt 2 notwendig. Stehen diese Ressourcen nicht im Einklang mit dem Aufwand, kann dies schnell die Belastbarkeit der entsprechend getroffenen Aussagen und Bewertungen infrage stellen. Ein wichtiger praktischer Aspekt für die Effizienz des gesamten Bewertungsprozesses ist dabei, wie oft es zu Abweichungen zwischen der Klassifikation des Leistungserbringerergebnisses in Schritt 1 und der anschließenden fach-

[33] Beispielsweise könnten die QI-Ergebnisse sich auf das Erfassungsjahr 2019 beziehen, die quantitative Auffälligkeitseinstufung jedoch sowohl auf die Daten des Erfassungsjahres 2018 als auch 2019. Der Qualitätsbericht der Krankenhäuser würde dann im Rahmen des Public Reportings nur die Ergebnisse für das Erfassungsjahr 2019 enthalten und das Ergebnis des Bewertungsprozesses. Aus Gründen der Transparenz müssten im Rückmeldebericht an den Leistungserbringer jedoch alle verwendeten Leistungserbringerergebnisse mitgeteilt werden.

lichen Bewertung in Schritt 2 kommt. Wird ein sehr starker statistischer Hinweis für die Auslösung des Stellungnahmeverfahrens gefordert, kann dies dazu führen, dass wichtige Qualitätsdefizite unentdeckt bleiben.

Die zentrale statistische Frage ist daher, wie genau und anhand welcher Informationen der in Kapitel 4 definierte Begriff „hinreichender Hinweis auf ein Qualitätsdefizit“ im Rahmen der Bewertungsart der statistischen Auswertungsmethodik operationalisiert werden soll, damit die verfügbaren Ressourcen zur fachlichen Bewertung effizient eingesetzt werden können.

Aus biometrischer Perspektive umfasst die Beantwortung dieser Frage gleichzeitig auch eine Beantwortung von Punkt 3.a des Auftrags, d. h. der „Optimierung der Diskriminationsfähigkeit der Indikatoren unter Prüfung verschiedener (auch derzeit noch nicht genutzter) statistischer Verfahren, Abwägung von Vor- und Nachteilen, z. B. in Anlehnung an das Verfahren „Planungsrelevante Qualitätsindikatoren“. Unter dem Begriff „Diskriminationsfähigkeit“ wird dabei die Klassifikationsgüte des Verfahrens zur quantitativen Auffälligkeitseinstufung verstanden. Um diese Frage im Rahmen der „Optimierung der Effizienz des Verfahrens“ diskutieren zu können, ist es notwendig, die quantitative Auffälligkeitseinstufung in einen formalen Rahmen einzubetten. Innerhalb eines solchen entscheidungstheoretischen Rahmens könnten verschiedene Methoden für die Operationalisierung des Begriffs „hinreichender Hinweis“ statistisch formuliert werden und deren „Effizienz“ mittels Verlustfunktionen verglichen werden. Die Formulierung der Methoden wird dabei in zwei Teile aufgeteilt:

- Teil 1: Optimierung der quantitativen Auffälligkeitseinstufung anhand der existierenden Datengrundlage (d. h. jährliche Erfassung) – siehe Abschnitt 5.3.
- Teil 2: Erweiterung von Teil 1, indem die Daten- bzw. Entscheidungsgrundlage auf einen 2-Jahres-Zeitraum erweitert wird – siehe Abschnitt 5.4. Der Grund für die Erweiterung ist, dass viele der prinzipiellen Festlegungen generisch für den 1-Jahres-Zeitraum diskutiert werden können. Die Erweiterung auf einen 2-Jahres-Zeitraum führt zu einer Reihe an zusätzlichen Festlegungen, die dann gesondert diskutiert werden.

5.2 Entscheidungstheoretische Modellierung des Bewertungsprozesses

Betrachtet wird die Situation für einen Qualitätsindikator (QI), der als Rate ausgedrückt werden kann und für den eine analytische Herangehensweise festgelegt wird. Es wird angenommen, dass der Referenzbereich für den QI $[0, R]$ ist, d. h. die im Indikator interessierenden Ereignisse sind unerwünschte Ereignisse, z. B. Todesfälle. Der Entscheidungsprozess der Bewertung eines Leistungserbringers für einen solchen Ratenindikator kann im Rahmen einer Spezifikationssprache für Entscheidungsprobleme als sogenanntes *Entscheidungsgraph* bzw. *Einflussdiagramm* (Englisch: *Influence Diagram*) grafisch dargestellt werden, vgl. z. B. Kapitel 9 in Jensen und Nielsen (2007). Entscheidungsgraphen erweitern die sogenannten *Bayesianischen Netzwerke* um Entscheidungen und illustrieren, welche Informationen für konkrete Entscheidungen vorliegen und welche Kriterien bzgl. der Entscheidungsfindung optimiert werden sollen.

5.2.1 Bayesianische Netzwerke und Einflussdiagramme

Bayesianische Netzwerke[34] sind statistische Modelle, welche probabilistische Abhängigkeitsstrukturen zwischen (Zufalls-)Variablen mittels gerichteter azyklischer Graphen illustrieren. Wie z. B. in Jensen und Nielsen (2007) oder Pearl (2009) erläutert, werden hierbei Variablen als runden Knoten im Netzwerk dargestellt, Pfeile symbolisieren bedingte Abhängigkeiten in der gemeinsamen Verteilung der Zufallsvariablen und das Fehlen von Pfeilen zwischen Variablen illustriert bedingte Unabhängigkeiten. Im Rahmen der analytischen Auswertung eines Ratenindikators können beispielsweise die Zusammenhänge zwischen der zugrunde liegenden Rate θ des Leistungserbringers (später auch Kompetenzparameter genannt), Fallzahl J und der Anzahl an unerwünschten Qualitätsereignissen O eines Leistungserbringers unter Berücksichtigung der innewohnenden Stochastizität wie in Abbildung 10 über Zufallsvariablen dargestellt werden:

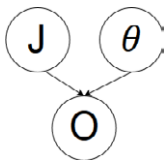

Abbildung 10: Bayesianisches Netzwerk zur Illustration der Zusammenhänge zwischen Fallzahl J, Kompetenzparameter ϑ und Anzahl unerwünschter Qualitätsereignisse O für einen Ratenindikator.

Zum Beispiel illustriert das Netzwerk in Abbildung 10, dass die Kenntnis von J und θ die Variable O probabilistisch festlegen und dass J und θ voneinander stochastisch unabhängig sind, d. h. die gemeinsame Wahrscheinlichkeitsverteilung $f(J,\theta,O) = f(O|J,\theta)f(J)f(\theta)$ ist. Die Darstellung alternativer Zusammenhänge, z. B. dass die Fallzahl auch die Kompetenz beeinflusst, ist im Diagramm über einen Pfeil von J zu θ möglich. Für alle Knoten ohne „Eltern“ im Graphen (d. h. Knoten, bei denen keine Pfeile in Richtung des Knotens laufen) muss im bayesianischen Netzwerk eine A-priori-Verteilung gewählt werden, welche das Vorwissen über die Zufallsgröße über eine Verteilung parametrisiert. Da die Wahl einer geeigneten A-priori-Verteilung ein wichtiger Aspekt der bayesianischen Inferenz darstellt, wird in Abschnitt 5.3.2.3 auf verschiedene Möglichkeiten der A-priori-Wahl eingegangen. Da die Fallzahl J in der Auswertungsmethodik, wie in Abschnitt 5.1.1 erläutert, als gegeben angesehen wird, und somit die A-priori-Verteilung für J entsprechend eine Punktmasse hat, muss nur für den Kompetenzparameter θ eine A-priori-Verteilung angenommen werden.

Wichtig ist an dieser Stelle, dass sich grafische Modelle dieser Art aus Sicht des IQTIG hervorragend dafür eignen, die komplexe Zusammenhangsstruktur zwischen den relevanten Variablen für die Auswertungsmethodik vereinfachend darzustellen. Grafische Modelle sind für diesen Zweck bereits bei der Beschreibung der statistischen Auswertungsmethodik bei Patientenbefragungen zum Einsatz gekommen (IQTIG 2018c, IQTIG 2018d). Das Einflussdiagramm stellt die entscheidungstheoretische Erweiterung des bayesianischen Netzwerkes dar. Es erweitert die Notation der bayesianischen Netzwerke mit zusätzlichen Symbolen für Nutzen- bzw. Verlustfunk-

[34] In der Literatur auch „Kausaldiagramm“ oder „Kausalgraph“ genannt (Pearl 2009).

tionen (Rauten), welche zu optimieren sind um zu Entscheidungen (Vierecke) zu gelangen. Für weitere Informationen zu Einflussdiagrammen siehe z. B. Jensen und Nielsen (2007).

5.2.2 Einflussdiagramm für den Bewertungsprozess

In Abbildung 11 stellt θ die zugrunde liegende Rate des Leistungserbringers für den betrachteten Indikator dar. Im analytischen Kontext wirkt sich dieser kausal auf das Behandlungsergebnis $O_j \in \{0,1\}$ jedes Falles aus. Dabei stellen die insgesamt J „Plättchen“ in der Abbildung die Wiederholung der Situation für jede Patientin bzw. jeden Patienten des Leistungserbringers da. Andere patientenseitige Faktoren (messbare bzw. erhobene Risikofaktoren x_j sowie nicht messbare bzw. nicht erhobene z_j) beeinflussen das Behandlungsergebnis zusätzlich, jedoch wird bei einem nicht risikoadjustierten Ratenindikator davon ausgegangen, dass diese Einflüsse marginal und für die Berechnung vernachlässigbar sind.

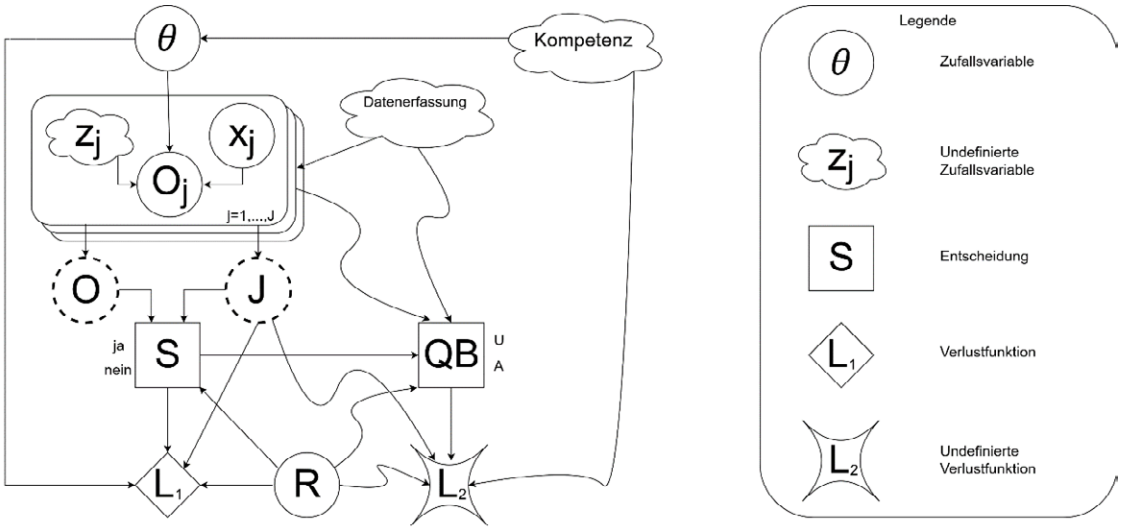

Abbildung 11: Schematisches Einflussdiagramm für das zweistufige Entscheidungsverfahren des Bewertungsprozesses für einen Ratenindikator. Die wichtigsten Komponenten sind in der Legende erklärt. Ein gerichteter Pfeil von einem Knoten A zu einem Knoten B bedeutet, dass B von A abhängt.

Im Rahmen eines statistischen Modells wird für die Variablen in der Abbildung angenommen, dass $O_j|\theta \sim \text{Bernoulli}(\theta)$, und dass für zwei Fälle j und j' des Leistungserbringers $O_j \perp O_{j'}|\,\theta$ gilt, d. h. dass – gegeben dem Kompetenzparameter – das Qualitätsergebnis beider Fälle voneinander statistisch unabhängig ist. Des Weiteren beschreibt $O = \sum_{j=1}^{J} O_j$ die Anzahl an Fällen mit dem für den Indikator interessierenden Ereignis in den J betrachteten Fällen. Das Qualitätsergebnis für jeden einzelnen Fall wird über die Zufallsvariable O_j dargestellt. Entsprechend ist die Gesamtzahl an unerwünschten Qualitätsereignissen die Summe der O_j . Somit liegt ein deterministischer Zusammengang zwischen den O_j und deren Summe O vor, der in der Abbildung dadurch hervorgehoben wird, dass die Zufallsvariable O als deterministische Variable (durch gestrichelte Umrandung) im Grafen dargestellt wird. Gleiches gilt für die Fallzahl J, welche direkt als Anzahl an „Plättchen“ im Grafen deterministisch bestimmt ist. Das obige Bernoulli-Modell

und die Annahme, dass die Behandlungsoutcomes – gegeben dem Kompetenzparameter – unabhängig voneinander sind, führt dazu, dass $O \sim \mathrm{Bin}(J, \theta)$. Das heißt, dass die Anzahl an interessierenden Ereignissen binomialverteilt ist mit Anzahlparameter J und zugrunde liegender Wahrscheinlichkeit[35] θ. Zur besseren Veranschaulichung der Rolle des Referenzwertes R wird dieser auch als Knoten im Diagramm repräsentiert, obwohl er keine Zufallsvariable darstellt – rein technisch entspricht die Verteilung einer Ein-Punkt-Verteilung mit der kompletten Wahrscheinlichkeitsmasse im Wert.

Die erste Entscheidung, die im Bewertungsprozess eines Leistungserbringers getroffen wird, ist die Beantwortung der Frage, ob anhand der beobachteten Daten für den Leistungserbringer ein Signal (S) in Form einer quantitativen Auffälligkeit für den betrachteten Indikator generiert werden soll. Anhand der vorliegenden Informationen soll also entschieden werden, ob man sich für $S = \text{ja}$ oder $S = \text{nein}$ entscheidet. Dies wird in Form einer sogenannten Strategie für die Entscheidung dargestellt, d. h. einer Funktion, die für jede vorhandene Information – hier ein Tripel aus beobachteter Anzahl an unerwünschten Qualitätsergebnissen, der Fallzahl und dem Referenzwert – eine entsprechende Entscheidungsalternative auswählt:

$$s\colon (o, J, R) \to \{\text{ja}, \text{nein}\}.$$

Das heißt, s stellt die Strategie der Auffälligkeitseinstufungsmethode da, wohingegen die Variable S das Ergebnis der Auffälligkeitseinstufung darstellt. Für die Entscheidung spielt der Referenzbereich $[0, R]$ des QI eine wichtige Rolle: gilt für den Leistungserbringer $\theta \leq R$, dann erfüllt der Leistungserbringer die QI-spezifischen Qualitätsanforderungen. Ist dagegen $\theta > R$, dann ist dies ein Hinweis auf ein mögliches Qualitätsdefizit. Die Schwierigkeit im analytischen Kontext ist jedoch, dass θ für den Leistungserbringer nicht bekannt ist, sondern nur anhand der beobachteten Daten J und o im inferenzstatistischen Sinne geschätzt werden kann, vgl. z. B. IQTIG (2016: Abschnitt 3.4) bzw. IQTIG (2019a: Abschnitt 13.1). Die Art und Weise, wie die beobachteten Daten unter Berücksichtigung von Unsicherheit Rückschlüsse auf θ erlauben, ist eine der Herausforderungen bei der Entwicklung der statistischen Auswertungsmethodik und spiegelt sich in der genauen Form der Funktion s wieder. Damit die Entscheidung, ob das Signal gegeben wird oder nicht, in einem Rahmen betrachtet wird, welcher es erlaubt, Optimalitätseigenschaften herzuleiten, wird die geeignete Wahl der Strategie s als Minimierung einer vordefinierten Verlustfunktion gesehen. Der formale Rahmen dafür wird im folgenden Abschnitt vorgestellt; in Abschnitt 5.3.1 wird das Konzept dann auf einen Ratenindikator für zwei Verlustfunktionen angewandt, um unterschiedliche Methoden für die Auffälligkeitseinstufung formal herzuleiten. Die Einbettung in diesen Rahmen erlaubt eine strukturierte Abwägung der Vor- und Nachteile beider Konzepte, welche auch qualitative Dimensionen und Aspekte wie Kommunizierbarkeit beinhalten.

Anhand des in Abbildung 11 dargestellten Entscheidungsdiagramms sind Strategien für die Entscheidungen bzgl. der quantitativen Auffälligkeit (Signal = ja oder nein) und der anschließenden fachlichen Prüfung zu entwickeln. Falls $s(o, J, R) = \text{nein}$, wird dies nicht weiter fachlich geprüft

[35] Im Rahmen der externen Qualitätssicherung werden solche Anteils- bzw. Quoten-Indikatoren auch gerne Ratenindikatoren genannt. Somit wird teilweise auch die Bezeichnung „zugrunde liegende Rate“ für den Parameter θ verwendet.

und die Qualitätsbewertung im zweiten Schritt entspricht automatisch dem Ergebnis der quantitativen Klassifikation. Falls $s(o, J, R) =$ ja, kommt es in einem zweiten Schritt zur fachlich-inhaltlichen Überprüfung des statistischen Hinweises. Dieser zweite Prozessschritt wird im Weiteren unter der Überschrift „Stellungnahmeverfahren" zusammengefasst. Die konkreten Schritte der Informationssuche für die Überprüfung sind stark indikatorabhängig, da sie die Prüfung von Einzelfällen, die z. B. nicht in den Rechenregeln abgebildete Ausnahmen darstellen (z. B. medizinische Ausnahmetatbestände) sowie eine Prüfung der Datenqualität beinhalten (vgl. Kapitel 4). Die genaue Abhängigkeits- und Informationsstruktur der Entscheidungsfindung im zweiten Schritt lässt sich daher nicht wirklich formalistisch im Rahmen eines Einflussdiagramms abbilden. Um dies zu verdeutlichen, werden die Knoten und Pfeile im Diagramm für diesen Teil bewusst vage als Wolken und geschwungene Pfeile dargestellt. Dies soll verdeutlichen, dass es sich um schwer formalisierbare Größen bzw. schwierig zu definierende Abhängigkeiten handelt. Wichtig im Diagramm ist aber die Feststellung, dass das Stellungnahmeverfahren sich auf zusätzliche Informationen stützt, die über die für den ersten Schritt der quantitativen Einstufung erfassten Daten hinausgehen. Diese zusätzliche Information wird im Diagramm über zusätzliche Pfeile in den Knoten, der die Entscheidung der Qualitätsbewertung (QB) repräsentiert, dargestellt.

Somit wird deutlich, dass die optimale Vorgehensweise sowohl vom Vorgehen der quantitativen Auffälligkeitseinstufung in Schritt 1 als auch des qualitativen Vorgehens in Schritt 2 abhängt. Ein zentraler Aspekt bei der Bewertung eines Qualitätsindikators pro Leistungserbringer ist dabei, inwieweit die Operationalisierung des Qualitätsaspekts durch den Indikator mit den vorliegenden Daten und Mengendefinitionen geeignet ist. Da jedoch sowohl der genaue Charakter der vorliegenden Informationen für die Entscheidung der Qualitätsbewertung als auch die entsprechende Verlustfunktion für die Entscheidung in Schritt 2 sich nur vage konkretisieren lassen, konzentriert sich die statistische Arbeit zur „Optimierung der Effizienz des Verfahrens" auf den quantitativen Schritt 1 – wohlwissend, dass die Entscheidung in diesem Schritt große Auswirkungen auf das Vorgehen in Schritt 2 hat. Diese Auswirkungen können jedoch nur qualitativ bzw. nur sehr indirekt quantitativ beschrieben werden. Da sich die formale Betrachtung, wie oben begründet, allein auf Schritt 1 der in Abbildung 11 dargestellten Entscheidungssequenz konzentriert, ist das Ziel, eine Strategie für die Entscheidung der quantitativen Auffälligkeit zu entwickeln, indem der erwartete Verlust bzgl. der Verlustfunktion L_1 minimiert wird.[36] Das heißt, das Interesse liegt in der Formulierung einer entsprechenden Verlustfunktion für Schritt 1, L_1, die das Entscheidungsproblem isoliert für Stufe 1 betrachtet und Informationen, Entscheidung und Verlustfunktion aus Schritt 2 werden im Folgenden nicht weiter betrachtet. Abbildung 12 zeigt das vereinfachte Entscheidungsdiagramm, welches nur die relevanten Teile für Schritt 1 enthält.

[36] Mit „erwartetem Verlust" ist der Erwartungswert der Verlustfunktion unter Verwendung einer spezifischen Entscheidungsstrategie gemeint. Der Erwartungswert entspricht einer gewichteten Summe über die Verluste der jeweiligen Entscheidung summiert über die verschiedenen Zustände des Entscheidungsproblems. Die Gewichte bei der Summierung entsprechen der jeweiligen Wahrscheinlichkeit für den Zustand – vergleiche z. B. Abschnitt 9.2 in Jensen und Nielsen (2007) oder gängige ökonometrische Literatur.

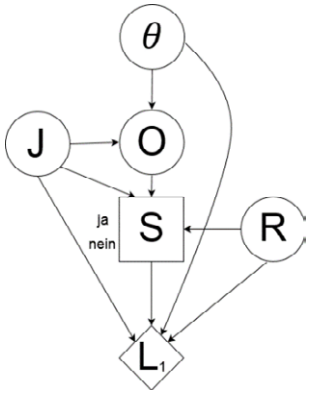

Abbildung 12: Vereinfachtes Entscheidungsdiagramm für die quantitative Auffälligkeitseinstufungsmethode bei einem Ratenindikator.

Tabelle 4: Vier-Felder-Schema zur Darstellung der Aufwände jeder Entscheidung bei der Klassifikation eines Leistungserbringers anhand dessen beobachteter Ergebnisse des Leistungserbringers, d. h. (o, J). Exemplarisch dargestellt ist die Situation, wenn nur Fehlklassifikationen zu Aufwänden führen.

		Wahrer Wert des LE	
		$\boldsymbol{\theta \leq R}$	$\boldsymbol{\theta > R}$
Ergebnis der quantitativen Auffälligkeitseinstufung	$s(o, J, R) = \text{nein}$	0	*fn*
	$s(o, J, R) = \text{ja}$	*fp*	0

Vereinfachend kann die Verlustfunktion auch als Vier-Felder-Schema dargestellt werden (siehe Tabelle 4). In Tabelle 4 wird die Situation dargestellt, wenn nur Fehlklassifikationen in Schritt 1 zu Aufwänden[37] führen. Dabei bezeichnet $fp > 0$ den Aufwand einer falsch-positiven Klassifikation (d. h. ein Leistungserbringer erfüllt den erwartbaren Qualitätsstandard und wird dennoch fälschlicherweise als quantitativ auffällig eingestuft, d. h. es liegt ein hinreichender Hinweis für ein Qualitätsdefizit vor, welcher dann mit entsprechendem Aufwand im Stellungnahmeverfahren korrigiert werden muss), $fn > 0$ die Aufwände einer falsch-negativen Klassifikation (d. h. ein Leistungserbringer erfüllt nicht den erwartbaren Qualitätsstandard, wird jedoch nicht als quantitativ auffällig eingestuft). Die Frage, in welcher Einheit dieser Aufwand zu messen ist, ist dabei nebensächlich; entscheidend ist dagegen das relative Verhältnis zwischen diesen Aufwänden. Jedoch ist zu beachten, dass Verlustfunktionen, die nur die Fehlklassifikationsaufwände berücksichtigen, den Aufwand für die Durchführung eines Stellungnahmeverfahrens ignorieren, weil diese immer bei der Entscheidung $s(o, J, R) = \text{ja}$ anfallen, d. h. auch wenn $\theta > R$. Des Weiteren berücksichtigt diese Art von Aufwandsbetrachtung auch nicht die Aufwände, welche durch ein mögliches Qualitätsdefizit verursacht werden, z. B. durch Schäden an Patienten und Patientinnen. Nur die Fehlklassifikationsaufwände zu berücksichtigen stellt also im Sinne von Abschnitt

[37] In diesem Kapitel werden die Begriffe „Aufwand“ bzw. „Aufwände“ verwendet, anstelle des gängigeren wirtschaftsmathematischen Begriffs „Kosten“, um zu verdeutlichen, dass die Betrachtungen weiter fassen, als monetäre Kosten.

5.2.3 eine starke Vereinfachung der Entscheidungssituation dar, erlaubt dagegen aber eine explizite Darstellung und Effizienzdiskussion.

5.2.3 Diskussion des entscheidungstheoretischen Ansatzes

Der in den vorigen Abschnitten vorgestellte Entscheidungstheoretische Ansatz formuliert explizit, dass es sich bei der quantitativen Auffälligkeitseinstufung um eine Klassifikation handelt. Die Wahl einer konkreten Vorgehensweise für die Auffälligkeitseinstufung fordert eine Abwägung zwischen Vor- und Nachteilen, die teilweise nur theoretisch betrachtet werden können, weil es keinen Goldstandard für die Klassifikation gibt. Um trotzdem die Abwägungen expliziter zu gestalten, werden Verlustfunktionen definiert, die dann unter bestimmten Annahmen zu optimieren sind, um z. B. kosteneffiziente Schwellenwerte für die Klassifikation von QI-Ergebnissen herzuleiten und somit auf theoretischer Grundlage eine Optimierung des Klassifikationsverfahrens herzuleiten. Die Schwierigkeit beim entscheidungstheoretischen Ansatz liegt darin, die Verlustfunktionen so zu formulieren, dass sie adäquat die vielen Aspekte der Entscheidung reflektieren. Dies erscheint im Rahmen der externen Qualitätssicherung extrem schwierig, da sehr viele unterschiedliche Interessen zu berücksichtigen sind. Als Beispiel: eine falsch-negative Klassifikation (d. h. die fälschliche Einstufung eines Leistungserbringer-Ergebnisses als „nicht auffällig“) hat beispielsweise zur Folge, dass kein Stellungnahmeverfahren stattfindet, was sich negativ auf die Behandlungsqualität zukünftiger Patientinnen und Patienten auswirken kann. Eine falsch-positive Klassifikation führt dagegen zu Aufwand auf Seiten des Leistungserbringers, der LAG bzw. Bundesstelle sowie möglicherweise der Patientinnen und Patienten, deren Entscheidung für einen Leistungserbringer durch Fehlinformationen beeinflusst wird. Diesen Aufwand überhaupt nur ansatzweise quantifizieren und gegeneinander abzuwägen ist keine mathematische, sondern vielmehr eine politische Entscheidung, ebenso wie jener, ob eine solche Aufwand-Nutzen-Abwägung überhaupt Grundlage für gesundheitspolitische Entscheidungen sein kann (Gerber-Grote et al. 2014).

Um trotzdem mithilfe des entscheidungstheoretischen Ansatzes zu einer formaleren Diskussion über Klassifikationsgüte und somit über Effizienzoptimierung zu kommen ist es hilfreich, lediglich auf eine wenige Aspekte der Entscheidung bei der Konstruktion der Verlustfunktion zu fokussieren. Dies könnte z. B. sein, nur auf die korrekte Klassifikation bzgl. des Kompetenzparameters zu fokussieren – egal, wie viele Patienten oder Patientinnen der Leistungserbringer behandelt und egal, wie groß die Abweichung vom Referenzwert ist. Alternativ können auch nur die patientenseitigen Nachteile einer falsch-negativen Klassifikationsentscheidung betrachtet und Entscheidungsregeln für die Klassifikation nur anhand von Grenzwerten für ein noch tolerierbares Ausmaß dieser Nachteile abgeleitet werden. Dies erlaubt zwar keine Abwägung der im Prozess entstehenden Gesamtaufwände für alle Parteien, hat jedoch den Vorteil, dass keine Aufwände für falsch-positive und falsch-negative Entscheidungen gegeneinander abgewogen werden müssen.

5.3 Methoden für die quantitative 1-Jahres-Auffälligkeitseinstufung

Ausgehend vom Rahmenkonzept in Abschnitt 5.1 und einer expliziten Formulierung der im vorherigen Abschnitt 5.2 vorgestellten Verlustfunktion L_1 werden im Folgenden drei Methoden der

quantitativen Auffälligkeitseinstufung vorgestellt, die unter unterschiedlichen Annahmen über die Verlustfunktion und die Herangehensweise im entscheidungstheoretischen Kontext jeweils die optimale Vorgehensweise zur quantitativen Auffälligkeitseinstufung darstellen. Zwei dieser Methoden finden im Rahmen der externen Qualitätssicherung bereits Anwendung: die rechnerische Auffälligkeit nach § 10 QSKH-RL sowie die statistische Auffälligkeit für die Verfahren im Rahmen der plan. QI-RL (IQTIG 2016), die im Folgenden *statistisch signifikante Auffälligkeit* genannt wird. Für die statistische signifikante Auffälligkeit wird eine bayesianische Alternative vorgeschlagen, die quasi identisch zur Methodik des Verfahrens für planungsrelevante Qualitätsindikatoren ist (siehe Abschnitt 5.3.1.2). Die bayesianische Variante bietet einen flexibleren Rahmen für quantitative Auffälligkeitseinstufungen und führt dabei zu ähnlichen Klassifikationsentscheidungen, wie die in der plan. QI-RL verwendete frequentistische Variante. Zusätzlich zur rechnerischen und statistisch signifikanten Auffälligkeit wird anschließend eine dritte Einstufungsmethodik, die *statistisch relevante Auffälligkeit*, vorgestellt, welche von anderen Aufwandsannahmen ausgeht als die statistisch signifikante Auffälligkeit und die rechnerische Auffälligkeit.

Im Folgenden wird von einem ratenbasierten Qualitätsindikator mit festem Referenzwert ausgegangen, der anhand der Daten eines Erfassungsjahres ausgewertet werden soll. Erweiterungen für risikoadjustierte Indikatoren und verteilungsabhängige Referenzwerte werden in den Abschnitten 5.3.2.2 und 5.3.4 diskutiert.

5.3.1 Lösung des Entscheidungsdiagramms

Die rechnerische, die statistisch signifikante und die statistisch relevante Auffälligkeitseinstufung werden im Folgenden definiert und die entsprechenden Annahmen erläutert, in der die jeweilige Auffälligkeitseinstufungs-Methode die optimale Lösung des Entscheidungsdiagramms darstellt. Dazu werden der Aufwand für jede Entscheidung als Funktion des Kompetenzparameters, des Referenzwertes und der binären Klassifikation wie in Abschnitt 5.2 beschrieben analysiert. Um ein Entscheidungsdiagramm zu lösen, d. h. um die Strategie zu finden, die entsprechend der Verlustfunktion optimale Entscheidungen trifft, muss sequentiell von hinten nach vorne jeweils die optimale Entscheidungsstrategie gefunden werden (Jensen und Nielsen 2007). Dabei wird die zugehörige Verlustfunktion nach dem vorgegebenen Zielkriterium minimiert. Im Folgenden wird als Zielkriterium die Minimierung des erwarteten Verlusts verwendet. Es können aber auch andere Zielkriterien wie z. B. Minimierung des Worst-Case-Ergebnisses o. Ä. von Relevanz sein.

Im Rahmen eines Ratenindikators ist die optimale Strategie für die Einstufung in Schritt 1 als folgendes Problem zu lösen

$$s(o,J,R) = \mathrm{argmin}_{s\in S}\mathrm{E}(L_1(s,J,R,\theta)) = \mathrm{argmin}_{s\in S}\int_0^1 L_1(s,J,R,\theta)\, f(\theta|o,J)d\theta,$$

wobei $S = \{\mathrm{ja}, \mathrm{nein}\}$ und der Erwartungswert über alle Zufallsgrößen in der Verlustfunktion gebildet wird, d. h. über alle zum Zeitpunkt der Entscheidung unbekannten Größen in der Verlustfunktion. Im konkreten Fall von Abbildung 11 wird nur über den Kompetenzparameter θ integriert, für welchen zur expliziten Berechnung des Integrals eine Wahrscheinlichkeits-Verteilung

$f(\theta|o,J)$ spezifiziert werden muss (siehe Abschnitt 5.3.2.1 bzw. Abschnitt 5.3.2.2). Da für die Wahl der optimalen Entscheidung bei der quantitativen Auffälligkeitseinstufung genau zwei Entscheidungsalternativen miteinander kontrastiert werden, gilt entsprechend, dass

$$s(o,J,R) = \mathrm{I}\{\mathrm{E}[L_1(s = \text{ja}, J, R, \theta)] < \mathrm{E}[L_1(s = \text{nein}, J, R, \theta)]\},$$

wobei $\mathrm{I}(\cdot)$ die Indikatorfunktion bezeichnet. Das heißt, dass die quantitative Auffälligkeitseinstufung so getroffen wird, dass der erwartete Verlust minimal ist. Hier wurde die Konvention gewählt, dass bei gleichem erwartetem Verlust der beiden Entscheidungsalternativen die Entscheidung – zugunsten des Leistungserbringers – für die Alternative „nein“ getroffen wird.

Um die Fallzahlabhängigkeit der gefundenen Entscheidungsregel explizit darzustellen, wird bei vorgegebenen J und Referenzwert R das kleinste o berichtet, welches zu einer Einstufung „quantiativ auffällig“ führt, d. h.

$$o_{\min}(J,R) = \mathrm{argmin}_{0 \le o \le J}\{s(o,J,R) = \text{ja}\}.$$

Eine Erkenntnis dabei ist, dass es in Abhängigkeit der Entscheidungsregel Fallzahlen geben kann, die so klein sind, dass es selbst bei der Beobachtung $o = J$ nicht zu einer quantitativen Auffälligkeit kommt. Dies bedeutet, dass es bei dieser Fallzahl *unabhängig vom Ergebnis* nicht genügend statistische Evidenz für ein mögliches Qualitätsdefizit geben kann. Die kleinste Fallzahl, ab der ein Leistungserbringer überhaupt als quantitativ Auffällig eingestuft werden kann, wird mit $J_{\min}(R) = \min\{J \ge 0 \mid s(o = J, J, R) = \text{ja}\}$ gekennzeichnet. Eine wichtige Frage für die Qualitätssicherung ist, wie mit Leistungserbringern umzugehen ist, die eine Fallzahl kleiner als $J_{\min}(R)$ haben. Dies wird in Abschnitt 5.3.1.5 diskutiert.

Ausgangspunkt für alle weiteren Betrachtungen ist die Formulierung unterschiedlicher Verlustannahmen analog zur im letzten Abschnitt vorgestellten Vier-Felder-Tafel (vgl. Tabelle 4). Während in der Literatur zu einschrittigen Klassifikationsverfahren üblicherweise kein Aufwand für richtig-positive Klassifikationen (d. h. in diesem Fall korrekt klassifizierten Qualitätsdefiziten) angenommen werden (vgl. u. a. Berger (2010)), ist diese Annahme für ein zweistufiges Klassifikationsverfahren, bestehend aus quantitativer Auffälligkeitseinstufung und fachlicher Klärung, nicht sinnvoll. Vielmehr ist es sinnvoll, sowohl im Falle einer richtig-positiven als auch einer falsch-positiven Klassifikation feste Aufwände ζ eines Stellungnahmeverfahrens anzunehmen. Betrachtet wird daher in Tabelle 5 eine im vgl. zu Tabelle 4 modifizierte Vier-Felder-Tafel.

Tabelle 5: Vier-Felder-Tafel der Aufwandsannahmen des zweischrittigen Klassifikationsproblems

		Wahrer Wert des LE	
		$\boldsymbol{\theta \le R}$	$\boldsymbol{\theta > R}$
Ergebnis der quantitativen Auffälligkeitseinstufung	$s(o,J,R) = \text{nein}$	0	*fn*
	$s(o,J,R) = \text{ja}$	ζ	ζ

Unterschiedliche Aufwandsannahmen für falsch-negative Klassifikationen (d. h. nicht entdeckte Qualitätsdefizite) führen zu unterschiedlichen Lösungen des Entscheidungsdiagramms bzw. zu unterschiedlichen optimalen Einstufungsverfahren. In Abschnitt 5.3.1.1 wird gezeigt, dass die

aus der QSKH-Richtlinie als rechnerische Auffälligkeit bekannte Einstufungsstrategie entscheidungstheoretisch optimal ist, unter der Annahme, dass die Schätzung des Kompetenzparameters θ keiner Unsicherheit unterliegt (enumerative Herangehensweise, also $\theta = o/J$,) und dass der Aufwand fn unabhängig von der Fallzahl J des Leistungserbringers und dem Ausmaß der Überschreitung $(\theta - R)$ des Referenzbereiches ist. Unter den gleichen Aufwandsannahmen, aber mit Berücksichtigung statistischer Unsicherheit (analytische Herangehensweise) wird in Abschnitt 5.3.1.2 gezeigt, dass die statistisch signifikante Einstufung die optimale Einstufungsstrategie darstellt. Schließlich wird in Abschnitt 5.3.1.3 betrachtet, welches die optimale Einstufungsstrategie ist, unter der Annahme, dass der Aufwand *fn* sowohl vom Ausmaß der Überschreitung des Referenzbereiches als auch von der Fallzahl des Leistungserbringers abhängt.

5.3.1.1 Rechnerische Auffälligkeit

Betrachtet wird nun ein ratenbasierter Qualitätsindikator mit o als beobachtete Anzahl an interessierenden Ereignissen, J als Fallzahl an behandelten Patientinnen und Patienten im betrachteten Zeitraum und R als Referenzwert des Indikators, d. h. mit Referenzbereich $[0, R]$. Die rechnerische Auffälligkeit nach § 10 QSKH-RL ist definiert als Entscheidungsstrategie

$$s_{\mathrm{rech}}(o, J, R) = I\left(\frac{\mathrm{o}}{\mathrm{J}} > R\right),$$

wobei $I(A)$ die Indikatorfunktion darstellt, welche 1 (bzw. „ja“) ist, falls das Boolesche Argument A erfüllt ist[38] und 0 (bzw. „nein“) ist, falls A nicht erfüllt ist.

Aus den in Tabelle 5 beschriebenen Aufwandsannahmen ergibt sich der erwartete Verlust $l_{\mathrm{rech}}(\cdot)$ für die quantitative Einstufung eines Leistungserbringerergebnisses für die beiden Entscheidungsmöglichkeiten $s_{\mathrm{rech}} \in \{\mathrm{ja}, \mathrm{nein}\}$ als jeweils

$$E\left[l_{\mathrm{rech}}\left(o, J, R|\theta = \frac{o}{J}, s = \mathrm{ja}\right)\right] = \zeta$$

für $s = \mathrm{ja}$ und

$$E\left[l_{\mathrm{rech}}\left(o, J, R|\theta = \frac{o}{J}, s = \mathrm{nein}\right)\right] = fn \cdot \mathrm{P}\left(\theta > R|\theta = \frac{o}{J}\right)$$

für $s = \mathrm{nein}$. Da in der enumerativen Herangehensweise keine Unsicherheit berücksichtigt wird, gilt

$$\mathrm{P}\left(\theta > R|\theta = \frac{o}{J}\right) = \mathrm{I}\left(\frac{o}{J} > R\right),$$

d. h. die Kompetenz des Leistungserbringers ist entweder sicher über dem Referenzwert $\mathrm{I}\left(\frac{o}{J} > R\right) = 1$ oder sicher unterhalb des Referenzwertes $\mathrm{I}\left(\frac{o}{J} > R\right) = 0$.

[38] im obigen Fall ist das Boolesche Argument A gleich $o/J > R$.

Die optimale Entscheidungsregel, welche den erwarteten Verlust minimiert, ist in diesem Fall daher:

$$s_{\text{rech}}\left(o, J, R|\theta = \frac{o}{j}\right) = \mathrm{I}\left(\zeta < fn \cdot \mathrm{I}\left(\frac{o}{J} > R\right)\right),$$

d. h.

$$s_{\text{rech}}\left(o, J, R|\theta = \frac{o}{J}\right) = \mathrm{I}\left(\frac{\zeta}{fn} < \mathrm{I}\left(\frac{o}{J} > R\right)\right).$$

Überwiegt der Aufwand eines Stellungnahmeverfahrens ζ dem Aufwand eines nicht verfolgten Qualitätsdefizites *fn*, d. h. im Fall

$$\frac{\zeta}{fn} \geq 1,$$

würde die optimale Einstufungsstrategie keine Leistungserbringer als quantitativ auffällig einstufen. Umgekehrt, falls

$$\frac{\zeta}{fn} < 1,$$

wird ein Leistungserbringer genau dann quantitativ auffällig wenn gilt

$$\frac{o}{J} > R.$$

Dies entspricht der rechnerischen Auffälligkeitseinstufung der QSKH-RL. Die rechnerische Auffälligkeit ist daher die optimale Entscheidungsstrategie zur Minimierung des erwarteten Verlusts, sofern angenommen wird, dass $\theta = o/J$ ist, $fn > \zeta$ ist und der Aufwand *fn* als unabhängig von Fallzahl und Ausmaß der Überschreitung des Referenzbereiches angenommen werden. Mit anderen Worten ist die rechnerische Auffälligkeitseinstufung nur optimal, wenn es keine Unsicherheit bei der Bestimmung des Kompetenzparameters gibt, d. h. im Rahmen der vorgestellten Taxonomie nur bei enumerativer Herangehensweise mit Vollerhebung. In allen anderen Situationen ist die Verwendung der rechnerischen Auffälligkeitseinstufung suboptimal.

5.3.1.2 Statistisch signifikante Auffälligkeit

Bei der statistisch signifikanten Auffälligkeit wird, anders als bei der rechnerischen Auffälligkeit, davon ausgegangen, dass die beobachtete Anzahl an interessierenden Ereignissen eine Zufallsvariable und der zugrunde liegende Parameter θ eines Leistungserbringers nicht direkt beobachtbar ist, sondern anhand der beobachteten Daten lediglich im inferenz-statistischen Sinne geschätzt werden muss. In der frequentistischen Version der statistischen Auffälligkeit wird anhand eines Hypothesentests zwischen der Nullhypothese $H_0: \theta \leq R$ und der Alternativhypothese $H_1: \theta > R$ entschieden, ob ein Leistungserbringerergebnis quantitativ auffällig oder unauffällig ist. In der bayesianischen Version wird diese Entscheidung über die A-posteriori-Wahrscheinlichkeit für die Nullhypothese unter zusätzlichen Verteilungsannahmen getroffen (siehe Abschnitte 5.3.2.1 und 5.3.2.2). Die frequentistische Version findet bereits Anwendung

im Verfahren nach plan. QI-RL (IQTIG 2016). Da die bayesianische Version jedoch zu fast identischen Klassifikationsstrategien und somit auch fast identischen Ergebnissen führt und gleichzeitig einen flexibleren Rahmen bietet, um auch Auffälligkeitseinstufungen bei komplexeren Indikatorarten wie z. B. Patientenbefragungen zu berücksichtigen, wird im Folgenden die bayesianische Version dargestellt und diskutiert.

Bayesianische Version der statistisch signifikanten Auffälligkeit

Ausgehend von den Aufwandsannahmen in der Vier-Felder-Tafel aus Tabelle 5 werden die gleichen Aufwandsannahmen wie für die rechnerische Auffälligkeitseinstufung gestellt. Dabei ist $fn > 0$ wieder der Aufwand einer falsch-negativen Klassifikation, $\zeta > 0$ der Aufwand eines Stellungnahmeverfahrens. Wieder wird davon ausgegangen, dass der Aufwand einer Fehlklassifikation des Kompetenzparameters immer gleich ist, d. h. nicht von Eigenschaften des Leistungserbringers, wie z. B. der Fallzahl und dem Ausmaß der Überschreitung des Referenzbereichs, abhängen. Zusätzlich wird, anders als bei der rechnerischen Auffälligkeit, davon ausgegangen, dass der Parameter θ nicht direkt beobachtbar ist, sondern nur aus den beobachteten Daten (und möglichen Vorinformationen) geschätzt werden kann. Diese Schätzung ist mit Unsicherheit behaftet, weshalb die Verteilung der Wahrscheinlichkeiten für den Parameter θ von Interesse ist. Für die Klassifikationsentscheidung ist dabei die Wahrscheinlichkeit der Nullhypothese $P(\theta \leq R)$, d. h. das Vorliegen erwartbarer Qualität, von Interesse. Aus dieser und den oben angenommenen Aufwänden für Fehlklassifikationen ergeben sich die beiden erwarteten Verluste als:

$$E\left[l_{\text{stat.sig.bayes}}(o, J, R, \theta, s = \text{ja})\right] = \zeta$$

und

$$E\left[l_{\text{stat.sig.bayes}}(o, J, R, \theta, s = \text{nein})\right] = fn \cdot \mathrm{P}(\theta > R).$$

Daraus ergibt sich

$$s_{\text{stat.sig.bayes}}(o, J, R, \alpha) = \quad \mathrm{I}\left(\zeta < fn \cdot (1 - \mathrm{P}(\theta \leq R))\right) = \quad \mathrm{I}\left(\mathrm{P}(\theta \leq R) < \frac{fn - \zeta}{fn}\right).$$

Bei der bayesianischen Version der statistisch signifikanten Auffälligkeit wird also die Wahrscheinlichkeit der Nullhypothese $\mathrm{P}(\theta \leq R)$ mit dem Aufwandsverhältnis $\frac{fn-\zeta}{fn}$ verglichen. Wird also die Entscheidungsregel $\mathrm{I}(\mathrm{P}(\theta \leq R) < \alpha)$ verwendet, dann ist diese Regel für die oben angenommene Verlustfunktion optimal, wenn der Schwellenwert $\alpha = (fn - \zeta)/fn$ gewählt wird. Die Wahl von α im Rahmen der (bayesianischen) statistisch signifikanten Auffälligkeit kann daher auch so interpretiert werden, dass über eine Festlegung des Aufwandsverhältnisses zwischen dem Aufwand eines Stellungnahmeverfahrens und dem Aufwand einer falsch-negativen Einstufung ein α festgelegt wird. Der Fall $\alpha = 0{,}05$ impliziert daher beispielsweise, dass das Ver-

hältnis des Aufwands eines Stellungnahmeverfahrens gegenüber einer falsch-negativen Klassifikation genau 19 zu 20 ist.[39] Mit dieser Wahl von $\alpha = 0{,}05$ werden somit dem Übersehen eines Qualitätsdefizites nur wenig mehr Aufwände zugeschrieben, als dem Aufwand eines Stellungnahmeverfahrens.[40] Wird der Aufwand für das Übersehen eines Qualitätsdefizits dagegen beispielsweise als doppelt so hoch bewertet wie der Aufwand eines Stellungnahmeverfahrens, so führt ein α von $(2\zeta - \zeta)/(2\zeta) = 0{,}5$ zur optimalen Entscheidungsregel, welches zu einer strengeren Bewertung als die rechnerische Auffälligkeit führen würde.

Die konkrete Berechnung der Wahrscheinlichkeit von $P(\theta \leq R)$ im Rahmen von ratenbasierten und risikoadjustierten Indikatoren wird in Abschnitt 5.3.2 erläutert. Insgesamt ist die Interpretation des Schwellenwertes α bei der Klassifikation im bayesianischen Rahmen sehr verwandt mit dem Signifikanzniveau α des einseitigen frequentistischen Tests – oft wird der Schwellenwert im obigen bayesianischen Verfahren daher auch einfach Signifikanzniveau genannt.

Im bayesianischen Kontext gilt $P(\theta \leq R) = 1 - P(\theta > R)$, d. h. obiges Vorgehen ist äquivalent zum Vorgehen, die Alternativhypothese $\theta > R$ anzunehmen, falls die Wahrscheinlichkeit dafür größer als $1 - \alpha$, d. h. z. B. größer als 95 % ist. Ist dies nicht der Fall, d. h. kommt es nicht zu einer Entscheidung für die Hypothese $\theta > R$, dann gilt im Umkehrschluss nicht unbedingt, dass ein hinreichender Hinweis für $\theta \leq R$ vorliegt (vgl. u. a. Altman und Bland (1995)). Soll zusätzlich entschieden werden, ob ein hinreichender Hinweis dafür vorliegt, dass ein Leistungserbringerergebnis im Referenzbereich liegt, könnte dies z. B. anhand der zusätzlichen Entscheidungsregel $I(\mathrm{P}(\theta \leq R) > 1 - \alpha))$ geschehen.

5.3.1.3 Statistisch relevante Auffälligkeit

Bisher wurde angenommen, dass der Aufwand für eine falsch-negative Einstufung eines Leistungserbringers fix ist und nicht von weiteren Faktoren abhängt, wie z. B. der behandelten Anzahl an Patienten und Patientinnen oder der Größe der Abweichung vom Referenzwert. Alternativ kann angenommen werden, dass die falsch-negative Klassifikation eines Leistungserbringers, der im nächsten Beobachtungszeitraum viele Fälle behandeln wird, aus Sicht der Patienten und Patientinnen höhere Aufwände verursacht, als jene eines Leistungserbringers mit wenigen Fällen. Gleiches gilt auch für die Größe der Abweichung des Kompetenzparameters vom Referenzwert bei zwei Leistungserbringern mit gleicher Fallzahl: Ist die Abweichung vom Referenzbereich sehr groß, dann ist die Anzahl an betroffenen Patienten und Patientinnen größer, als bei einem Leistungserbringer mit gleicher Fallzahl, aber mit geringer Abweichung vom Referenzbereich. In beiden Fällen liegt die Annahme zugrunde, dass das Stellungnahmeverfahren zu einer Aufdeckung von möglichen Qualitätsdefiziten führt und damit einer Verbesserung der Behandlungsqualität für alle nachfolgend behandelten Patientinnen und Patienten nach sich zieht. Die „Aufdeckung“ eines Qualitätsdefizits bei Leistungserbringern mit

[39] Verwendet man stattdessen die Verlustfunktion aus Tabelle 4 in Abschnitt 5.2.2, welche nur die Fehlklassifikationsaufwände betrachtet und die Aufwände des Stellungnahmeverfahrens ignoriert, dann ist das optimale α gleich $fn/(fn + fp)$.

[40] Siehe auch Abschnitt 5.3.1.5 zu Limitationen der entscheidungstheoretischen Verlustbetrachtung.

großer Fallzahl wäre in dieser Betrachtungsweise wichtiger als die „Aufdeckung“ eines Qualitätsdefizits bei Leistungserbringer mit kleiner Fallzahl.

Eine alternative Operationalisierung des *hinreichenden Hinweises auf ein Qualitätsdefizit* in dieser Betrachtungsweise beruht nicht allein auf der statistische Evidenz für die Nullhypothese bzgl. des Kompetenzparameters des Leistungserbringers, d. h. $H_0: \theta \leq R$, sondern auf der *Relevanz* dieser Abweichung für Patienten und Patientinnen. Speziell hängt dies von der Größe der Abweichung vom Referenzwert sowie der Anzahl an betroffenen Patienten und Patientinnen beim Leistungserbringer in der Zukunft ab. Selbst wenn als Referenzrahmen nur die Fallzahl im *nächsten* Bewertungszeitraum, d. h. die Fallzahl im nächsten Erfassungsjahr als Metrik genommen wird, ist diese zum Zeitpunkt der Auffälligkeitsentscheidung unbekannt. Vereinfachend wird daher im Folgenden angenommen, dass die Fallzahl J im aktuellen Beobachtungszeitraum ein guter Schätzer für die Fallzahl im nächsten Beobachtungszeitraum ist. Die mit einer Fehlklassifikation verbundenen Aufwände aus patientenzentrierter Sichtweise lassen sich dann wie in untenstehender Matrix als abgewandelte Version der Vier-Felder-Tafel in Tabelle 5 quantifizieren:

Tabelle 6: Aufwandsannahmen für die statistisch relevante Einstufungsmethode

		Wahrer Wert des LE	
		$\boldsymbol{\theta \leq R}$	$\boldsymbol{\theta > R}$
Ergebnis der quantitativen Auffälligkeitseinstufung	$s_{\text{stat.rel}}(o, J, R) = \text{nein}$	0	$fn \cdot (\theta - R)_+ \cdot J$
	$s_{\text{stat.rel}}(o, J, R) = \text{ja}$	ζ	ζ

In dieser Darstellung ist dabei $(\theta - R)_+ \cdot J$ die über die von Referenzwert R hinaus tolerierte Anzahl an Fällen mit unerwünschtem Ereignis, mit $(\theta - R)_+ := \max(0, \theta - R)$. Des Weiteren ist $fn > 0$ der mit jedem dieser Fälle assoziierte Aufwand, und $\zeta > 0$ wiederum der Aufwand eines Stellungnahmeverfahrens. Vereinfachend kann die Verlustfunktion auch als $fn \cdot (\theta - R)_+ \cdot J$ bei $s_{\text{stat.rel}} = \text{nein}$ und ζ bei $s_{\text{stat.rel}} = \text{ja}$ dargestellt werden.

Beispiel: Hat ein Leistungserbringer in einem Indikator mit $R = 15\%$ im betrachteten Erfassungsjahr $J = 40$ Fälle bei einer zugrunde liegenden Rate von $\theta = 20\%$ behandelt, d. h. der Kompetenzparameter ist 20 %, so liegt die erwartete Anzahl an Fällen mit unerwünschtem Qualitätsergebnis genau $0{,}05 \cdot 40 = 2$ über der vom Referenzwert tolerierten Anzahl (in diesem Fall $0{,}15 \cdot 40 = 6$). Die konkrete Entscheidung, ob dieser Leistungserbringer als quantitativ auffällig eingestuft werden soll, hängt dann wiederum vom Aufwandsverhältnis zwischen ζ und fn ab.

Die optimale Entscheidungsregel lässt sich aus den beiden Verlustfunktionen

$$E[l_{\text{stat.rel}}(o, J, R, \theta, s = \text{ja})] = \zeta$$

und

$$E[l_{\text{stat.rel}}(o, J, R, \theta, s = \text{nein})] = \int fn \cdot (\theta - R)_+ \cdot J \cdot f(\theta | o, J)\, d\theta$$

herleiten als Entscheidungsregel

$$s_{\text{stat.rel}}(o, J, R, \theta) = \quad \mathrm{I}\left(\zeta < \int fn \cdot (\theta - R)_+ \cdot J \cdot f(\theta|o, J)\, d\theta\right),$$

mit $f(\theta|o, J)$ als Dichtefunktion der Wahrscheinlichkeitsverteilung von θ (gegeben die beobachteten Daten o, J). Das heißt, es wird der Aufwand für die Durchführung eines Stellungnahmeverfahrens ζ dem erwarteten Aufwand für die über die vom Referenzwert hinaus tolerierte Anzahl an Patienten mit unerwünschtem Qualitätsergebnis gegenübergestellt. Im Folgenden wird dabei vereinfachend $fn = 1$ gesetzt, weil die optimale Auswertungsstrategie lediglich vom Aufwandsverhältnis von ζ/fn abhängt. Als Beispiel: $\zeta = 1$ bedeutet, dass der Aufwand eines Stellungnahmeverfahrens gleichzusetzen sind mit dem Aufwand für die vom Referenzwert tolerierte Anzahl an Fällen mit unerwünschtem Behandlungsergebnis plus genau einem weiteren Fall. Wird die Grenze für quantitative Auffälligkeit als $\zeta = 1$ gesetzt, werden daher alle Leistungserbringer quantitativ auffällig, bei denen mindestens ein Fall mit unerwünschtem Behandlungsergebnis über die vom Referenzwert tolerierte Anzahl hinaus erwartet wird.

Zur expliziten Berechnung der statistisch relevanten Auffälligkeitseinstufung muss die Wahrscheinlichkeitsverteilung von θ (gegeben die beobachteten Daten o, J) spezifiziert werden. Dies erfolgt in einem bayesianischen Rahmen und wird am Beispiel des Beta-Binomial-Modelles in Abschnitt 5.3.2 für einen Ratenindikator diskutiert.

5.3.1.4 Zusammenfassung: Annahmen zu Aufwänden und Herangehensweise der drei Einstufungsverfahren

Die in Abschnitt 5.3.1.1 bis 5.3.1.3 vorgestellten Einstufungsverfahren werden tabellarisch hinsichtlich der zu Grunde liegenden Annahmen gegenübergestellt:

Tabelle 7: Übersichtstabelle über die Annahmen und Entscheidungsregeln der drei vorgestellten quantitativen Auffälligkeitseinstufungsmethoden

Auffälligkeitseinstufungsmethode	rechnerisch	statistisch signifikant	statistisch relevant
Herangehensweise	enumerativ	analytisch	analytisch
Aufwand für Stellungnahmeverfahren	ζ	ζ	ζ
Aufwand für falsch-negative Klassifikation	fn	fn	$fn \cdot (\theta - R)_+ \cdot J$
Entscheidungsregel	$I\left(\frac{o}{J} > R\right)$, (sofern $\frac{fn}{\zeta} < 1$)	$I\left(\mathrm{P}(\theta \leq R) < \frac{fn - \zeta}{fn}\right)$	$\mathrm{I}\left(\zeta < \int fn \cdot (\theta - R)_+ \cdot J \cdot f(\theta\|o, J)\, d\theta\right)$

5.3.1.5 Limitationen der entscheidungstheoretischen Verlustfunktionenbetrachtung

Die Messung und Steigerung der Effizienz des Strukturierten Dialogs, wie im Auftragstext dieses Berichtes formuliert (G-BA 2018a), erfordert es, Aufwand und Nutzen des Strukturierten Dialoges bzw. des Stellungnahmeverfahrens in expliziter Form miteinander ins Verhältnis zu setzen. Die Vorgehensweise, Strategien zur quantitativen Auffälligkeitseinstufung aus verschiedenen expliziten Aufwandsannahmen herzuleiten, bietet eine strukturierte Möglichkeit, die Annahmen des jeweiligen Einstufungsverfahrens transparent zu machen. Dabei ist zu berücksichtigen, dass bestimmte Vereinfachungen für die Formulierung der Verlustfunktionen getroffen wurden: Beispielsweise hängt das beobachtete Indikatorergebnis eines Leistungserbringers in der Regel nicht nur von der zugrunde liegenden Kompetenz ab, sondern auch von weiteren Faktoren wie beispielsweise der Qualität der Datenerfassung in den QI-relevanten Feldern oder unbeobachteten patientenseitigen Risikofaktoren (vgl. das Entscheidungsdiagramm in Abbildung 11). Wenn davon auszugehen ist, dass diese Faktoren nicht vollständig bei der Konstruktion und Messung eines Indikators berücksichtigt werden (können), kann dies dazu führen, dass ein strengerer oder weniger strenger Schwellenwert zu einer besseren Entscheidungsregel führt als die oben hergeleitete, welche diese Störfaktoren nicht berücksichtigt. Gibt es in einem Indikator beispielsweise unbeobachtete patientenseitige Risikofaktoren, die zu einer erhöhten Wahrscheinlichkeit der im Indikator betrachteten interessierenden Ereignisse führen, und werden diese nicht über eine Risikoadjustierung oder in der Festlegung des Referenzbereichs berücksichtigt, so führt dies in den Beobachtungen der Zufallsvariable O zu einer sogenannten Überdispersion, vgl. z. B. Spiegelhalter (2005). Damit wird die Entscheidung ob $\theta > R$ vorliegt unsicherer, d. h. beispielsweise, dass die Wahrscheinlichkeit der Nullhypothese $P(\theta \leq R)$ ohne Berücksichtigung der Störfaktoren tendenziell überschätzt wird. Somit führt bei der statistisch signifikanten Auffälligkeitseinstufungsmethode ein kleineres α als jenes aus der Aufwandsabwägung zur optimalen Entscheidungsregel. Dies gilt unter der Annahme, dass diese unbeobachteten patientenseitigen Faktoren bei der fachlichen Bewertung bewertet und zu einer Entlastung des Leistungserbringers führen. Analoge Überlegungen treffen auf die statistisch relevante Auffälligkeitseinstufung zu. Dies sollte bei der Festlegung des Signifikanzniveaus bzw. des Schwellenwertes ζ für die statistisch relevante Einstufung immer Berücksichtigung finden.

Eine weitere Vereinfachung ist, bei der Auslösung eines Stellungnahmeverfahrens die gleichen Aufwände anzunehmen, unabhängig davon, ob sich im Stellungnahmeverfahren ein Qualitätsdefizit bestätigt oder nicht. Prinzipiell könnten auch hier die Aufwände einer falsch-positiven Klassifikation höher sein, als bei einer richtig-positiven Klassifikation, z. B. $\zeta + fp$ vs. ζ mit einem weiteren Aufwandsparameter fp. Jedoch ist das gewählte Vorgehen $fp = 0$ zu setzen auch dadurch zu begründen, dass das anschließende Stellungnahmeverfahren im 2. Schritt eine falsch-positive Klassifikation verhindern würde. Eine weitere zusätzliche Vereinfachung, die getroffen wurde, ist es, die Sensitivität und Spezifität des Stellungnahmeverfahrens im 2. Schritt des Entscheidungsproblems unberücksichtigt zu lassen: Auch im Stellungnahmeverfahren können Fehler bei der Einstufung passieren, die wiederum unterschiedliche Aufwände verursachen können.

Grundsätzlich fordert die Aufwand-Nutzen-Betrachtung der vorangegangenen Abschnitte eine explizite Festlegung des Aufwandsverhältnisses von Stellungnahmeverfahren gegenüber nicht entdeckten Qualitätsdefiziten. Dabei ist zu beachten, dass verschiedene Akteure sehr unterschiedliche Aufwandsperspektiven auf den Bewertungsprozess haben. Diese Komplexität der Aufwand-Nutzen-Betrachtung ist mit den vereinfachenden Annahmen der vorangegangenen mathematischen Herleitungen nur bedingt vereinbar. Trotzdem bietet der entscheidungstheoretische Rahmen eine Offenlegung der Abwägungen und kann eine Stütze bieten, um die Tuning-Parameter[41] der jeweiligen Einstufungsverfahren zu interpretieren. Prinzipiell legen die Schwellenwerte fest, ab wann ein hinreichender Hinweis auf ein Qualitätsdefizit vorliegt. Eine zu klärende Frage ist, ob dieses Niveau einheitlich für alle QS-Verfahren, Qualitätsindikatoren und Verwendungszwecke festgelegt werden soll, oder situativ anzupassen ist. Beispielsweise kann es Gründe geben, dass man für manche Qualitätsindikatoren (z. B. Sterblichkeitsindikatoren) weniger statistische Evidenz verlangt, bevor man im Rahmen eines Stellungnahmeverfahrens den Kompetenzparameter des Leistungserbringers klären möchte, als bei anderen Qualitätsindikatoren. Auch können die Konsequenzen des Verwendungszweckes (Bsp. plan. QI-RL) dazu führen, dass man ein sehr hohes Niveau ansetzen möchte, dann aber im zweiten Schritt eine sehr umfangreiche fachliche Bewertung ansetzt, die entsprechend transparent dokumentiert ist, um den gravierenden Konsequenzen der Bewertung gerecht zu werden. Alle diese Gründe lassen sich im Rahmen der Verlustfunktion begründen. Eine weitere Diskussion zu diesem Thema und konkrete Empfehlungen findet sich in Abschnitt 5.6.

5.3.2 Bayesianische Modellierung

Die im vorangegangenen Abschnitt vorgestellten Ansätze zur Lösung des Entscheidungsdiagramms (Abbildung 11), müssen je nach Indikatorart mit konkreten Wahrscheinlichkeitsmodellen zur Modellierung der Verteilung des Kompetenzparameters θ und der beobachteten Behandlungsergebnisse O_j ausgeführt werden. In der Literatur vorherrschend sind dabei bayesianische Modelle (vgl. u. a. Liu et al. (2003), George et al. (2017), Ash et al. (2012), Christiansen und Morris (1996)), welche die Wahrscheinlichkeitsverteilung der beobachteten Behandlungsergebnisse O_j für den j'ten Patient oder Patientin bei gegebener Kompetenz θ auf obersten Ebene (Patientenebene) und die Wahrscheinlichkeitsverteilung von θ in einer zweiten Ebene (Leistungserbringerebene) modellieren. Diese Art der hierarchischen Modellierung berücksichtigt die Cluster-Struktur der erhobenen Daten, in der mehrere Patientinnen und Patienten jeweils von einem Leistungserbringer behandelt werden. Kernidee dieser Modelle ist, dass man durch die Modellierung einer Verteilung von θ Vorwissen modelliert, welches zusammen mit den beobachteten Daten in die Schätzung der Kompetenz des Leistungserbringers einfließt. Vorwissen kann dabei aus verschiedenen Quellen kommen:

- der Historie eines Leistungserbringers
- Expertenwissen
- Daten anderer Leistungserbringer

[41] Als Tuning-Parameter sind gemeint: das Signifikanzniveau α in der statistisch signifikanten Einstufungsmethode bzw. der Schwellenwert ζ in der statistisch relevanten Einstufungsmethode

Dieses Vorwissen wird in der bayesianischen Statistik als sog. *A-priori-Verteilung* bezeichnet. Liegt kein oder nur sehr wenig Vorwissen vor, so spricht man von einer sog. *vagen* A-priori-Verteilung. Unter Beobachtung der Daten eines Leistungserbringers wird dieses Vorwissen mit der durch die Daten gewonnen Evidenz aufdatiert, und man erhält die sog. *A-posteriori-Verteilung*. Das Verwenden von bayesianischen Methoden erlaubt es, Vorwissen und Informationen transparent und zu unterschiedlichem Grad in die Einstufung von Leistungserbringern einzubeziehen, und somit je nach Zweck zu optimieren. Bei sogenannten *Empirical-Bayes-Ansätzen* werden A-priori-Informationen aus den Ergebnissen der Leistungserbringer geschätzt, was insbesondere bei verteilungsabhängigen Einstufungen (d. h. mit Perzentilreferenzbereich) sinnvoll ist. Je nach Indikatorart können A-priori-Verteilung und Annahmen über das Zustandekommen von Behandlungsergebnissen flexibel modelliert und, beispielsweise im Falle von Patientenbefragungen, sogar über mehr als zwei zu berücksichtigende Ebenen modelliert werden. Erweiterungen von George et al. (2017) beschreiben, wie durch explizite Modellierung der Fallzahlabhängigkeit Zusammenhänge zwischen Fallzahl und Ergebnisqualität bei der Einstufung berücksichtigt werden können. Dies führt zu sog. fallzahlabhängigen *Shrinkage Targets*, die es erlauben, je nach Fallzahl eine andere A-priori-Verteilung anzunehmen.

Im Folgenden werden zwei gängige Modelle für anteilsbasierte Indikatoren und für risikoadjustierte Indikatoren ($\frac{O}{E}$) vorgestellt.

5.3.2.1 Anteilsbasierte Indikatoren

Im Falle von anteilsbasierten Indikatoren werden patientenseitige Einflüsse x_j (beobachtete) und z_j (unbeobachtete) auf das Behandlungsergebnis O_j einer Patientin oder eines Patienten $j = 1, \dots, J$ nicht bei der Inferenz des Kompetenzparameters θ berücksichtigt. Die Wahrscheinlichkeit

$$\pi_j = P(O_j = 1|\theta)$$

für das Auftreten des Behandlungsergebnisses $O_j = 1$ wird daher innerhalb eines Leistungserbringers als für alle Patientinnen und Patienten gleich angenommen und kann direkt mit dem Kompetenzparameter θ gleichgesetzt werden: $\pi_j = \theta$ für alle $j = 1, \dots, J$. Die Summe $O = \sum_{j=1}^{J} O_j$ folgt somit einer Binomial-Verteilung mit Erfolgswahrscheinlichkeit θ und J wiederholungen.

Das Beta-Binomial-Modell

Die Festlegung der A-priori-Verteilung von θ als Beta-Verteilung $\mathrm{Beta}(\alpha, \beta)$ mit Parametern $\alpha, \beta > 0$ führt zum sog. Beta-Binomial-Modell (vgl. z. B. (Carlin und Louis)):

$$O \sim \mathrm{Bin}(J, \theta),$$

und

$$\theta \sim \mathrm{Beta}(\alpha, \beta).$$

Die Beta-Verteilung kann je nach Wahl der sog. A-priori-Parameter α und β sehr verschiedene Formen annehmen, und bietet somit einen sehr flexiblen Rahmen, um Vorwissen wiederzugeben. Der A-priori-Erwartungswert der Beta-Verteilung lautet

$$\mathbf{E}(\theta|\alpha,\beta) = \frac{\alpha}{\alpha+\beta}.$$

Die A-priori-Varianz der Beta-Verteilung

$$\mathbf{Var}(\theta|\alpha,\beta) = \frac{\alpha\cdot\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}.$$

Die Familie der Beta-Verteilungen wird dabei gewählt, weil es sich um eine sog. *konjugierte A-prior-Verteilung* zur Binomial-Verteilung handelt. Diese besitzt die besondere Eigenschaft, dass unter Beobachtung von J Behandlungsfällen mit $O = o$ interessierenden Ereignissen, die A-posteriori-Verteilung von θ erneut einer Beta-Verteilung folgt. Dabei lässt sich zeigen (Carlin und Louis 2008), dass

$$\theta|o,J \sim \mathrm{Beta}(\alpha+o, \beta+J-o).$$

Der A-posteriori-Erwartungswert dieser Verteilung lautet

$$\mathbf{E}(\theta|o,J,\alpha,\beta) = \frac{\alpha+o}{\alpha+\beta+J}.$$

Diese Formulierung bietet eine Interpretation der Parameter α und β als a priori angenommene Anzahl von Fällen mit interessierendem Ereignis (α) und Fällen ohne interessierendem Ereignis (β) an.

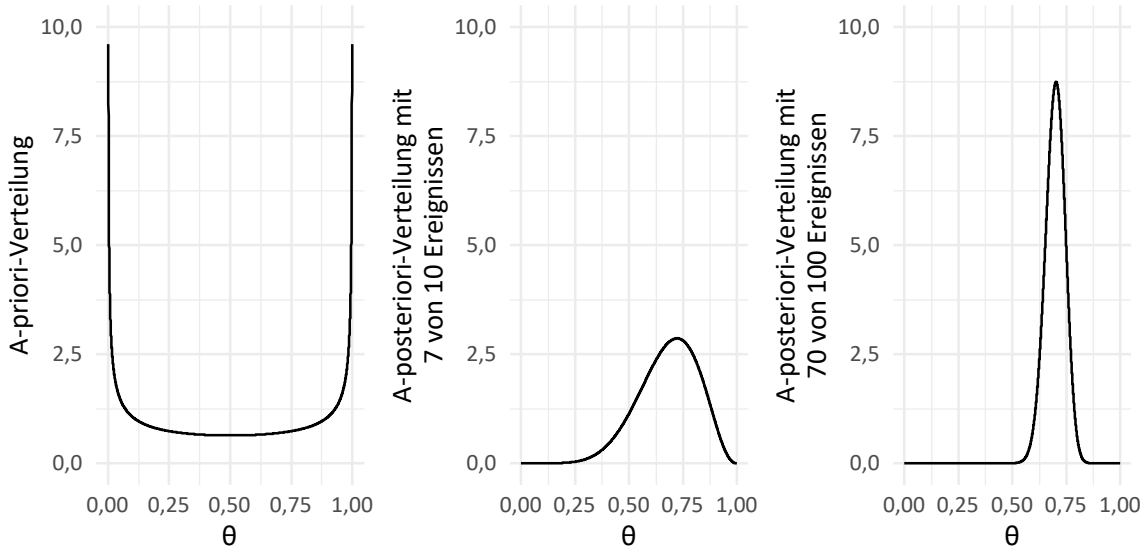

Abbildung 13: links: Betaverteilung mit Parametern $a = b = ½$; Mitte: resultierende A-posteriori-Verteilung bei 7 von 10 Ereignissen; rechts: resultierende A-posteriori-Verteilung bei 70 von 100 Ereignissen

Abbildung 13 veranschaulicht die A-priori- und A-posteriori-Verteilung von θ. Die linke Grafik zeigt dabei die A-priori-Verteilung $\mathrm{Beta}(\frac{1}{2},\frac{1}{2})$. Die mittlere und die rechte Grafik zeigen, wie sich

das Aufdatieren der A-priori-Verteilung mittels der Informationen aus den beobachteten Daten auswirkt. Beide Grafiken zeigen die A-posteriori-Verteilungen im Fall einer beobachteten Ereigniswahrscheinlichkeit von $o/J = 0{,}7$. Der Unterschied zwischen der mittleren und der rechten Grafik entsteht durch die Anzahl an beobachteten Fällen: In die mittlere Grafik gehen nur 10 Fälle ein, in die rechte Grafik dagegen 100. Dadurch erhalten im ersten Fall die Daten mehr und die A-priori-Informationen über den Parameter entsprechend weniger Gewicht als im zweiten Fall.

Alternativ lässt sich der A-posteriori-Erwartungswert auch als gewichtetes Mittel von A-priori-Erwartungswert und roher Rate $\frac{o}{J}$ ausdrücken:

$$\mathbf{E}(\theta|o, J, \alpha, \beta) = \rho \cdot \frac{o}{J} + (1 - \rho) \cdot \frac{\alpha}{\alpha + \beta}.$$

Dabei ist das Gewicht ρ, $0 \leq \rho \leq 1$, für die Mittelung wie folgt

$$\rho = \frac{J}{J + \alpha + \beta}.$$

Je kleiner J gegenüber $\alpha + \beta$ ist, desto geringer wird die Evidenz der beobachteten Rate $\frac{o}{J}$ gegenüber dem A-priori-Erwartungswert gewichtet. Dieses Phänomen wird auch als *Shrinkage* (dt. „Schrumpfung“) bezeichnet, weil die beobachtete Rate zum A-priori-Erwartungswert schrumpft. ρ wird auch als *Shrinkage-Gewicht* bezeichnet (vgl. u. a. Morris (1983)).

Beispiel für Einstufungsverfahren mit Beta-Binomial-Annahme

Der Referenzwert und die Ausrichtung eines Indikators definieren einen fallzahlunabhängigen Referenzbereich, bei ratenbasierten Indikatoren je nach Ausrichtung entweder $[0, R]$ oder $[R, 1]$, der angibt, welcher Wertebereich für den zugrunde liegenden Parameter θ erwartbare Qualität wiederspiegelt. Der Schwellenwert $o_{\min}(J, R)$ einer vorgegebenen Entscheidungsregel $s(o, J, R)$ ist dagegen definiert als der Wert, ab dem ein beobachtetes Leistungserbringerergebnis mit der Entscheidungsregel als quantitativ auffällig eingestuft wird. Während der Referenzbereich von der quantitativen Auswertungsmethodik unabhängig ist, hängen die Schwellenwerte für quantitative Auffälligkeit direkt von der verwendeten Klassifikationsmethodik für die Bewertung und deren Tuning-Parameter ab und sind, wie oben hergeleitet, in der Regel fallzahlabhängig. Die Fallzahlabhängigkeit der Schwellenwerte ist eine natürliche Konsequenz der Unsicherheit, die entsteht, wenn anhand der beobachteten QI-Ergebnisse Rückschlüsse auf den Kompetenzparameter gezogen werden. Aufgrund der Fallzahlabhängigkeit werden Schwellenwerte oft in sogenannten Funnelplots (IQTIG 2016, Spiegelhalter et al. 2012) dargestellt. Über diese Darstellungsweise werden im Folgenden die Schwellenwerte $o_{\min}(J, R)$ für die drei oben hergeleiteten Klassifikationsmethoden verglichen, d. h. für die rechnerische, statistisch signifikante und statistisch relevante quantitative Auffälligkeitseinstufung. Der Vergleich wird anhand eines Qualitätsindikators mit Referenzbereich $[0{,}10\%]$ durchgeführt. Eine interaktive Version dieser Darstellung, bei der sowohl der Referenzbereich als auch die Tuning-Parameter der Klassifikationsmethoden (wie z. B. α oder ζ) varriiert werden können, ist als Shiny-App (Chang et al. 2019) unter https://iqtig.shinyapps.io/funnelplot zu finden.

Abbildung 14 zeigt einen Funnelplot für den oben beschriebenen Ratenindikator mit Referenzbereich $[0,\ 10\%]$ und drei verschiedenen Methoden zur quantitativen Auffälligkeitseinstufung: rechnerische Auffälligkeit, statistisch signifikante Auffälligkeit (mit $\alpha = 5\%$) und statistisch relevante Auffälligkeit (mit $\zeta = 2$).

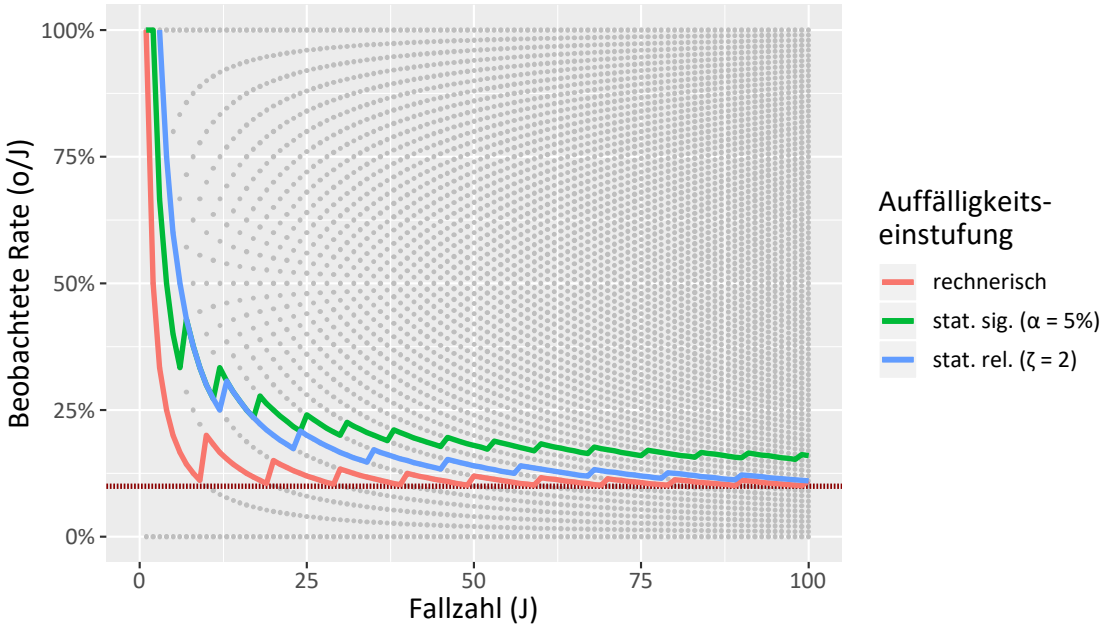

Abbildung 14: Funnelplot für die quantitative Auffälligkeitseinstufung der drei Klassifikationsmethoden für einen Indikator mit Referenzbereich [0–10 %].

Für eine gegebene Fallzahl J sind dabei nur beobachtete Raten $0, 1/J, 2/J, \ldots, 1$ möglich, die als Punkte in der Grafik dargestellt werden. Liegt ein beobachtetes Leistungserbringerergebnis auf oder oberhalb der entsprechenden Linie für den Schwellenwert der Methode, wird dieser Leistungserbringer als quantitativ auffällig in dem entsprechenden Indikator klassifiziert. Aus der Abbildung wird deutlich, dass sowohl die rechnerische als auch die statistisch signifikante Auffälligkeit mit steigender Fallzahl gegen den Referenzwert R konvergieren. Die Schwellenwerte der rechnerischen Auffälligkeit liegen dabei durchgängig unter jenen der statistisch signifikanten Auffälligkeit; dies ist aber nicht grundsätzlich so, sondern von der Wahl von α abhängig: wie oben beschrieben sind die Schwellenwerte bei einem α von $0{,}465$ nahezu identisch zu denen der rechnerischen Auffälligkeit.

Die Schwellenwerte der statistisch relevanten Auffälligkeit konvergieren scheinbar auch gegen den Referenzwert R; dies ist aber nur scheinbar der Fall. Für Fallzahlen $J > 240$ (in der Abbildung nicht dargestellt) liegen die Schwellenwerte dieser Einstufungsmethodik knapp unter dem Referenzwert, z. B. wird ein Ergebnis bei $J = 241$ mit $o = 24$ d. h. mit $o/J = 9{,}9585\ \%$ bereits auffällig, bei $J = 5000$ ab 9,6 %. Das bedeutet, dass bei dieser Klassifikationsmethodik grundsätzlich auch Leistungserbringerergebnisse als quantitativ auffällig eingestuft werden können, die prinzipiell nicht rechnerisch auffällig sind. Dies liegt darin begründet, dass über die A-posteriori-Dichte integriert wird und daher der Fall eintreten kann, dass der Erwartungswert der Verlustfunktion für ein o größer ist als ζ, obwohl für dieses o der Punktschätzer o/J innerhalb des

Referenzbereichs liegt. Wenn dies keine gewünschte Eigenschaft der Klassifikationsmethodik für die quantitative Auffälligkeit ist, könnte beispielsweise die rechnerische Auffälligkeit zu einer zusätzlichen Bedingung für die quantitative Auffälligkeit gemacht werden.

Die statistisch relevante Auffälligkeit ist bei dem betrachteten Indikator mit $R = 10\,\%$ bis zu einer Fallzahl von $J = 11$ gleich oder weniger streng als die statistisch signifikante Auffälligkeit. Ab dieser Fallzahl sinken die Schwellenwerte der statistisch relevanten Auffälligkeit jedoch schneller als jene der statistisch signifikanten Auffälligkeit, d. h. für diese Fallzahlen führt die statistisch relevante Auffälligkeit zu strengeren Schwellenwerten. Der Grund hierfür liegt in der Wahl der Verlustfunktion: Für die statistisch signifikante Auffälligkeitseinstufung relevant ist lediglich eine möglichst sichere Klassifikation des Kompetenzparameters des Leistungserbringers, d. h. ob $\theta \leq R$ oder $\theta > R$, egal, wie groß eine potentielle Abweichung vom Referenzwert ist bzw. wie viele Patienten und Patientinnen davon betroffen sind. Die statistisch relevante Einstufung gewichtet dagegen die Relevanz der Abweichung für Patienten und Patientinnen höher. Beide Verfahren nehmen an, dass die Qualität der Leistungserbringer nicht fallzahlabhängig ist.

Tabelle 8 verdeutlicht die unterschiedlichen Schwellenwerte für ausgewählte Fallzahlen.

Tabelle 8: Kritische Zahl interessierender Ereignisse, ab der Leistungserbringer in den drei Auffälligkeitseinstufungsmethoden als quantitativ auffällig bewertet werden.

	Schwellenwert $o_{\min}(J, R)$ der		
Fallzahl J	**s_{rech}**	**$s_{stat.sig.bayes}$ ($\alpha = 5\,\%$)**	**$s_{stat.rel}$ ($\zeta = 2$)**
1	1/1 = 100 %	1/1 = 100 %	-
2	1/2 = 50 %	2/2 = 100 %	-
3	1/3 = 33 %	2/3 = 67 %	3/3 = 100 %
5	1/5 = 20 %	2/5 = 40 %	3/5 = 60 %
10	2/10 = 20 %	3/10 = 30 %	3/10 = 30 %
20	3/20 = 15 %	5/20 = 25 %	4/20 = 20 %
50	6/50 = 12 %	9/50 = 18 %	7/50 = 50 %
100	11/100 = 11 %	16/100 = 16 %	11/100 = 11 %

Tabelle 8 verdeutlicht, dass die statistisch relevante Auffälligkeitseinstufung für Leistungserbringer mit kleinen Fallzahlen am liberalsten ist; Leistungserbringer mit weniger als 3 behandelten Fällen können unabhängig von ihrem beobachteten Ergebnis nicht quantitativ auffällig werden. Ab einer Fallzahl von $J = 10$ ist die statistisch relevante Auffälligkeit dann gleich streng oder strenger als die statistisch signifikante Auffälligkeit. Für eine Abwägung der Vor- und Nachteile der verschiedenen Methoden siehe Abschnitt 5.6.

5.3.2.2 Risikoadjustierte Indikatoren

Ausgehend vom Entscheidungsdiagramm in Abbildung 11 sind in der quantitativen Auffälligkeitseinstufung risikoadjustierter Qualitätsindikatoren bei der Inferenz des Kompetenzparameters θ auch explizit gemessene Einflüsse x_j der Patientinnen und Patienten $j = 1, \dots, J$ auf die jeweiligen Behandlungsergebnisse O_j zu berücksichtigen. Anders als im Falle anteilsbasierter (nicht risikoadjustierter) Indikatoren muss die Wahrscheinlichkeit

$$\pi_j = P(O_j = 1 | x_j, \theta)$$

für das Eintreten des Behandlungsergebnisses $O_j = 1$ als einerseits abhängig von der Kompetenz des Leistungserbringers und andererseits als abhängig von patientenseitigen Einflussgrößen x_j modelliert werden. Somit ist die Summe $O = \sum_{j=1}^{J} O_j$ anders als im Falle anteilsbasierter Indikatoren, wegen der unterschiedlichen Wahrscheinlichkeiten, nicht mehr binomial-verteilt, sondern als generalisiert-binomial-verteilte Zufallsvariable

$$O \sim \mathrm{GBin}(\pi_1, \dots, \pi_L) \tag{1}$$

zu betrachten. Das IQTIG verwendet zur Risikoadjustierung in der Regel das in der Epidemiologie unter der Bezeichnung „standardisierte Mortalitätsrate“ bzw. „Morbiditätsrate (SMR)“ bekannte Verhältnis von beobachteter zu erwarteter Anzahl interessierender Ereignisse (o/e) (im Folgenden immer als *die SMR* bezeichnet).[42] Dies entspricht einer indirekten Standardisierung (Keiding und Clayton 2014), bei der die beobachtete Zahl an interessierenden Ereignissen $O = o$ einer erwarteten Zahl an interessierenden Ereignissen *e* gegenübergestellt wird. Dabei wird die erwartete Zahl an interessierenden Ereignissen e in der Regel als Summe von Wahrscheinlichkeiten

$$e = \sum_{j=1}^{J} e_j$$

mit $e_j = P(O_j = 1 | x_j)$ berechnet, welche meist mithilfe logistischer Regressionsmodelle prognostiziert werden. Dabei fließen nur patientenseitige Einflussgrößen x_j, nicht aber die Kompetenz des Leistungserbringers θ in die Modellierung von e_j bzw. e ein. Im Allgemeinen gilt also $e_j \neq \pi_j$. Im Weiteren werden die Wahrscheinlichkeiten e_j vereinfachend als feste gegebene Größen betrachtet, deren Schätzunsicherheit vernachlässigt werden kann.[43]

[42] Äquivalent ist es das Verhältnis von beobachteter zu erwarteter *Rate* an interessierenden Ereignissen zu betrachten, d. h. $\frac{o/J}{e/J}$.

[43] Die Validität dieser Vereinfachung ist auf Ebene einzelner Qualitätsindikatoren zu prüfen. Insbesondere für Risikoadjustierungsmodelle mit kleiner Datenbasis kann die Schätzunsicherheit im Allgemeinen nicht vernachlässigt werden, und erfordert Erweiterungen der hier vorgestellten Methodik. Desweiteren ist davon auszugehen, dass nicht erhobene bzw. nicht in der Modellierung berücksichtigte Risikofaktoren z_j zur Unsicherheit der Schätzung von e_j beitragen.

Das Poisson-Gamma-Modell

Das Poisson-Gamma-Modell (vgl. u. a. Liu et al. (2003), Christiansen und Morris (1996)) modelliert die Verteilung der beobachteten Anzahl an interessierenden Ereignissen O als Poisson-verteilte Zufallsvariable

$$O \sim \mathrm{Po}(\theta \cdot e).$$

Dabei ist θ die für einen Leistungserbringer zu schätzende SMR und e die erwartete Anzahl an interessierenden Ereignissen, basierend auf dem Risikoadjustierungsmodell für die Patientenrisiken (also ohne Berücksichtigung des Leistungserbringereinflusses). Die Annahme, dass man O als poisson-verteilt modellieren kann, und nicht generalisiert binomial-verteilt wie in Gleichung (1), wird durch die sog. Poisson-Approximation begründet: Ist das Verhältnis von interessierenden Ereignissen und Fallzahl klein, so kann die Verteilung einer generalisiert binomial-verteilten Zufallsvariable durch eine Poisson-Verteilung approximiert werden.[44]

Im Poisson-Gamma-Modell wird für θ dabei eine Gamma-A-priori-Verteilung angenommen

$$\theta \sim \mathrm{Gamma}\left(\alpha, \frac{\alpha}{\mu}\right),$$

mit $\mu > 0, \alpha > 0$. Der Parameter μ ist als A-priori-Erwartungswert von θ zu interpretieren, d. h. $E(\theta) = \mu$. Die A-priori-Varianz von θ ist

$$\mathrm{Var}(\theta) = \frac{\mu^2}{\alpha},$$

wodurch der Parameter α als Präzisionsparameter interpretiert werden kann: Je kleiner α, desto größer ist die A-priori-Unsicherheit in der Verteilung von θ. Abbildung REF zeigt die Wahrscheinlichkeitsdichte einer Gamma-Verteilung für verschiedene Werte des Parameters α und $\mu = 1$.

[44] Maßgeblich für die Validität dieser Approximation ist dabei, dass die für den Approximationsfehler geltende obere Schranke $2\sum_{j=1}^{J} \pi_j^2$ klein ist. Diese Annahme kann auch in der Praxis bei Indikatoren mit häufigen Ereignissen und kleinen Fallzahlen verletzt sein, weshalb im Einzelfall komplexere Modelle nötig werden können.

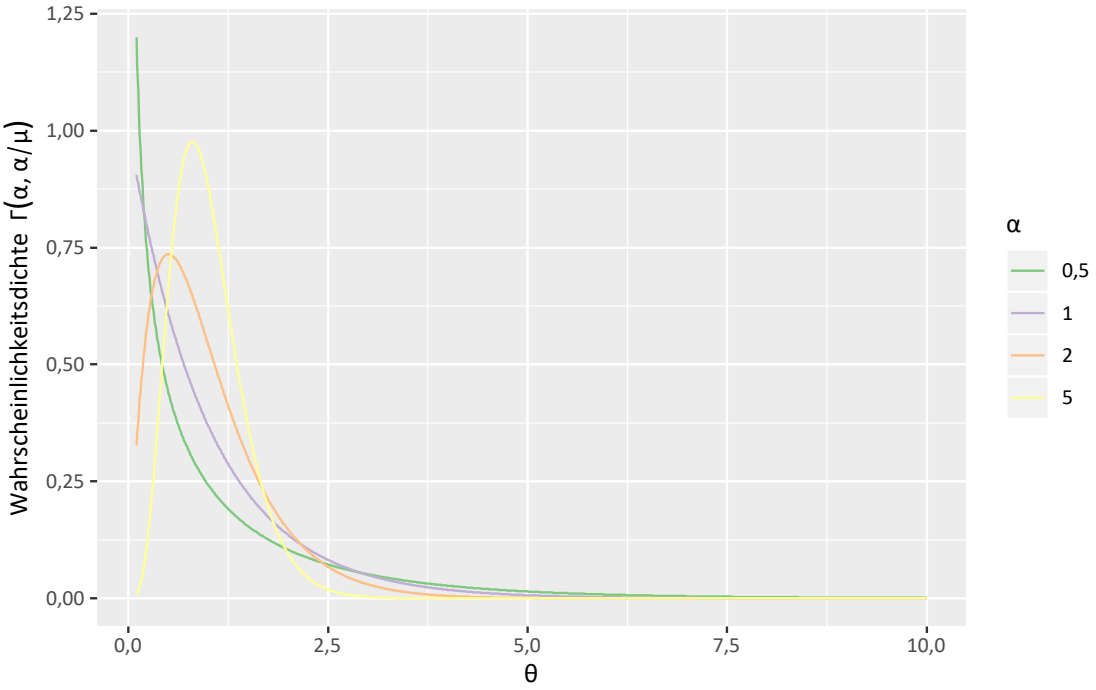

Abbildung 15: Wahrscheinlichkeitsdichte einer Gamma-Verteilung mit ϑ = 1 bei verschiedenen A-priori-Werten von α

Zu erkennen ist, dass sich die meiste Wahrscheinlichkeitsmasse im Bereich von ca. 0 bis 2 konzentriert. Während die Verteilung für $\alpha = 1/2$ eine Polstelle bei 0 besitzt, wird mit $\alpha = 1$ dem Wert $\theta = 0$ die höchste Wahrscheinlichkeit beigemessen. Für $\alpha = 2$ liegt der Modus der Verteilung bei $\theta = 1/2$. Mit wachsendem α konzentriert sich die Wahrscheinlichkeitsverteilung beim Modus $\mu \cdot \frac{\alpha-1}{\alpha}$, der für große Werte von α nahe am Erwartungswert μ liegt.

Unter Beobachtung einer Anzahl $O = o$ interessierender Ereignisse und erwarteter Anzahl e ist die A-posteriori-Verteilung von θ (Christiansen und Morris 1996):

$$\theta|o,e \sim \text{Gamma}\left(\alpha + o, e + \frac{\alpha}{\mu}\right).$$

Der Erwartungswert dieser A-posteriori-Verteilung ist:

$$\mathbf{E}(\theta|o,e) = \frac{o+\alpha}{e+\alpha/\mu}.$$

Dies legt eine weitere Interpretation der Parameter α und μ nahe: α entspricht einer beobachteten Anzahl von interessierenden Fällen, die man a priori annimmt und zu den tatsächlichen beobachteten interessierenden Fällen hinzuaddiert. Analog ist $\frac{\alpha}{\mu}$ die a priori erwartete Anzahl von interessierenden Ereignissen, die man zur erwarteten Anzahl e hinzuaddiert.

Um den Einfluss der A-priori-Verteilung auf das Shrinkage-Verhalten zu verstehen, kann der A-posteriori-Erwartungswert wiederum als gewichtetes Mittel des SMR-Schätzers o/e und des A-priori-Erwartungswerts μ ausgedrückt werden:

$$\mu' := \mathbf{E}(\theta|o,e) = \rho \cdot \frac{o}{e} + (1-\rho) \cdot \mu,$$

mit $\rho = \frac{e \cdot \mu}{\alpha + e \cdot \mu}$. Je größer α im Verhältnis zu $e \cdot \mu$ gewählt wird, desto stärker ist der Einfluss der A-priori-Verteilung, d. h. der Shrinkage-Effekt wächst mit α. Bei fester Wahl von α und μ steigt der Einfluss der beobachteten Daten mit steigendem e und der Shrinkage-Effekt nimmt ab.

Die A-posteriori-Varianz lautet

$$\mathrm{Var}(\theta|o,e) = \mu' \frac{\rho}{e}.$$

Im Kontext von Empirical-Bayes-Methoden, welche die A-priori-Parameter anhand empirischer Daten aller Leistungserbringer bestimmen, wird das Shrinkage-Gewicht ρ auch als Reliabilität der Messung interpretiert (Adams 2009), denn es gilt:

$$\rho = \frac{\mathrm{Var}(\theta|\alpha,\mu)}{\mathrm{Var}(\theta|\alpha,\mu) + \mathrm{Var}\left(\frac{O}{e}|\alpha,\theta=\mu\right)}.$$

Somit kann der Shrinkage-Faktor ρ auch als Varianzen-Verhältnis der sog. *between-provider-variance* $\mathrm{Var}(\theta|\alpha,\mu)$ und der Summe der Varianzen von $\mathrm{Var}(\theta|\alpha,\mu)$ und $\mathrm{Var}\left(\frac{O}{e}|\alpha,\theta=\mu\right)$ (*within-provider-variance*) interpretiert werden. Der A-posteriori-Erwartungswert μ' wird aus diesem Grund auch als *reliability adjusted estimate* bezeichnet (vgl. u. a. Dimick et al. (2010), Krell et al. (2014b), Krell et al. (2014a)).

Beispiel für Einstufungsverfahren mit Poisson-Gamma-Annahme

In Abbildung 16 werden verschiedene Einstufungsstrategien anhand eines Funnelplots gegenüber gestellt.

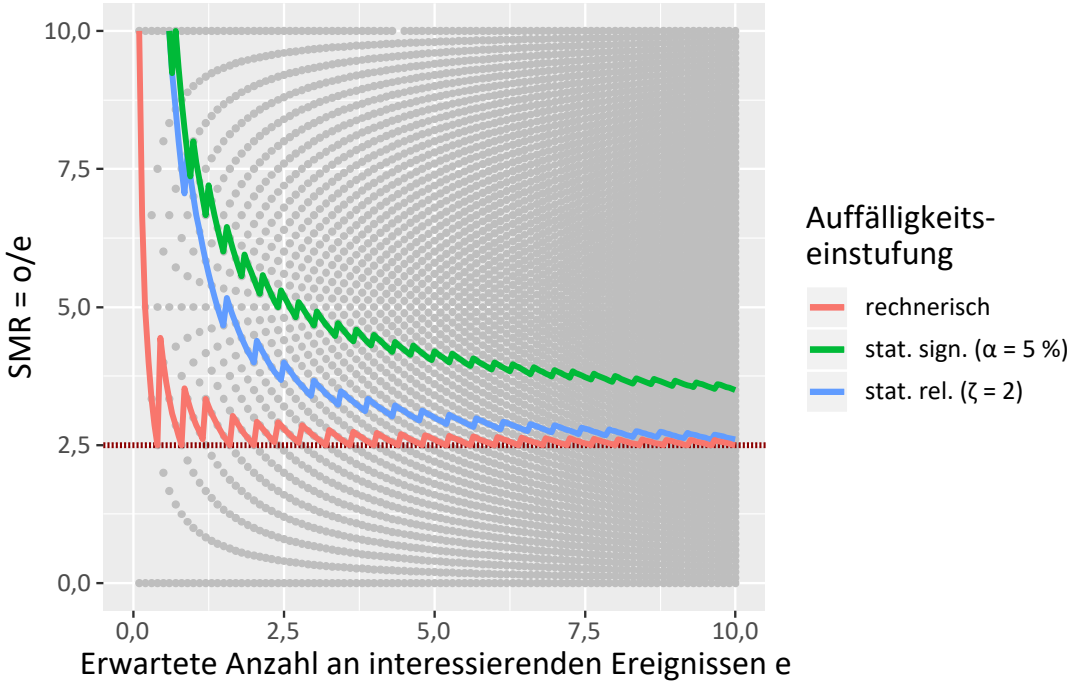

Abbildung 16: Funnelplot-Darstellung verschiedener Auffälligkeitseinstufungsmethoden für risikoadjustierte Qualitätsindikatoren.

Auf der x-Achse wird die Anzahl interessierender Ereignisse *e* dargestellt. Auf der y-Achse werden SMR-Werte dargestellt. Graue Punkte in der Grafik stellen hypothetische Ergebnisse dar, die von Leistungserbringern angenommen werden können. Diese folgen je nach beobachteter Anzahl an interessierenden Ereignissen *o* den Hyperbel-Ästen $\frac{1}{e}, \frac{2}{e}, \frac{3}{e}, \ldots$ usw. Jede der farbigen Linien entspricht einer Einstufungsstrategie. Die dunkelrote horizontale Linie stellt einen Referenzwert von $R = 2{,}5$ dar, welcher der rechnerischen Auffälligkeitseinstufung entspricht. Blau sind die Schwellenwerte eingezeichnet, ab welcher Leistungserbringer bei der Einstufung nach statistischer Relevanz für $\zeta = 2$ quantitativ auffällig werden würden. Dabei wurde eine vage A-prioi-Verteilung mit $\alpha = 1/2$ und $\mu = 1$ gewählt; zum Vergleich in Grün der Schwellenwert aus der statistisch signifikanten Auffälligkeitseinstufung (α = 5 %). Im Vergleich zur rechnerischen Auffälligkeit führen sowohl die Einstufung nach statistischer Relevanz als auch die Einstufung nach statistischer Signifikanz zu weniger strengen Schwellenwerten, vor allem im Bereich kleiner Fallzahlen. Je größer der Wert ζ gewählt wird, desto großzügiger fällt die Einstufung insbesondere bei niedrigen Werten von *e* aus. Gegenüber der statistisch signifikanten Auffälligkeitseinstufung fällt die Einstufung für große Werte von *e* bei der statistisch relevanten Auffälligkeitseinstufung strenger aus, während für kleine Werte von *e* die statistisch signifikante Auffälligkeitseinstufung strenger ist (vgl. auch Tabelle 9).

Die folgende Tabelle zeigt die Anzahl interessierender Ereignisse *o*, die je nach Einstufungsstrategie und erwarteter Zahl *e* bei einer A-priori-Verteilung von $\alpha = 1/2$ und $\mu = 1$ nötig ist.

Tabelle 9: Kritische Anzahl an interessierenden Ereignissen, die zur quantitativen Auffälligkeit führt, je nach erwarteter Anzahl an interessierenden Ereignissen und Einstufungsstrategie

erwartete Anzahl an interessierenden Ereignissen	**kritische Anzahl beobachteter Ereignisse**			
e	**rechnerische Auffälligkeit**	**statistisch relevant (ζ = 1)**	**statistisch relevant (ζ = 2)**	**statistisch signifikant (α =0,05)**
0,1	1	7	13	4
0,5	2	4	6	6
1,0	3	5	7	8
2,0	5	7	8	11
5,0	13	13	15	21
10,0	25	24	26	35

5.3.2.3 Weitere Indikatorarten

In der externen Qualitätssicherung kommen über die ratenbasierten Indikatoren und die risikoadjustierten Indikatoren hinaus eine Reihe weiterer Indikatorarten bei der Berechnungsart zum Einsatz. Beispiele sind der Geburtshilfe-Index, der e/N-Indikator in HTXM-MKU, die Follow-up-

Indikatoren in den Transplantationsbereichen sowie die Follow-up-Indikatoren bei HEP, KEP und HSM (Anlage 3 QSKH-RL Follow-up-Indikatoren). Auch diese lassen sich einfach im bayesianischen Rahmen darstellen, falls nicht – wie beim e/N-Indikator – bereits bayesianisch formuliert. Auch die statistische Auswertungsmethodik für Qualitätsindikatoren bei Patientenbefragungen ist bereits mittels bayesianischen Modellen formuliert (IQTIG 2018c, IQTIG 2018d) und passt so nahtlos in das statistische Rahmenkonzept, sobald diese im Rahmen der DeQS-RL auszuwerten sind.

Die in IQTIG (2017c) und IQTIG (2018b) sowie Hengelbrock und Höhle (2019) beschriebene statistische Auswertungsmethodik für Follow-up-Indikatoren besteht für Ereigniszeit-adjustierte Ratenindikatoren bereits aus einer bayesianischen Vorgehensweise, welche Zensierung und Trunkierung bei der Datenerhebung berücksichtigt. Die Berechnung der risikoadjustierten, Ereigniszeit-adjustierten Indikatoren, welche als Kompetenzparameter das standardisierten Inzidenzratenverhältnis haben, lässt sich bayesianisch im Rahmen des oben beschriebenen Poisson-Gamma-Modells interpretieren. Insofern sollte die entwickelte Follow-up-Methodik prinzipiell auch für Sozialdaten-gestützte Follow-up-Indikatoren vorbereitet sein, wobei es sich erst in der Praxis und bei Vorliegen von Echtdaten zeigen wird, ob die längeren Zeitverzüge durch die Lieferung der Sozialdaten bei diesen Indikatoren weitere Modifikationen bei der statistischen Auswertungsmethodik erfordern.

5.3.2.4 Wahl der A-priori-Verteilung in bayesianischen Modellen

Der bayesianische Ansatz zur Modellierung der Wahrscheinlichkeitsverteilungen für die Kompetenzparameter bietet einen flexiblen Weg, unterschiedliche Indikatorarten und Verlustannahmen in einer einheitlichen Systematik zu behandeln. Gleichzeitig erzwingt der bayesianische Ansatz mit der expliziten Modellierung der A-priori-Verteilung Transparenz über Vorwissen und Annahmen herzustellen. Die A-priori-Verteilung kann dabei wesentlichen Einfluss auf die Einstufungsmethodik haben. Im Beta-Binomial-Modell ist die Summe $\alpha + \beta$ (d. h. die Anzahl virtuell angenommener A-priori-Beobachtungen) entscheidend dafür, wie stark die A-priori-Verteilung gegenüber den beobachteten Daten ist. Im Poisson-Gamma-Modell ist es der Parameter α, welcher die Stärke des A-priori-Einflusses misst.

Wenn Shrinkage-Effekte unerwünscht sind, so müssen vage A-priori-Verteilungen mit niedrigen Werten von $\alpha + \beta$ (Beta-Binomial-Fall) bzw. α (Poisson-Gamma-Fall) gewählt werden. Ein Spezialfall der vagen A-priori-Verteilung ist mit $\alpha = \beta = 1/2$ (Beta-Binomialfall) der sog. Jeffreys-Prior. Dieser spiegelt statistisch gesehen maximale Unwissenheit wieder. Wegen $\alpha + \beta = 1$ ist gewährleistet, dass das Gewicht des Vorwissens nie stärkeres Gewicht als das der beobachteten Daten besitzt. Man kann zeigen, dass die statistisch signifikante Einstufung mit Jeffreys-Prior nahezu äquivalent zur statistisch signifikanten Einstufung im Verfahren *Planungsrelevante Qualitätsindikatoren* mit mid-P-Werten ist (Brown et al. 2001)

Neben der Stärke der A-priori-Verteilung muss auch das Shrinkage-Target, d. h. der A-priori-Erwartungswert so modelliert werden, dass er den erwarteten Anforderungen entspricht. Beispielsweise könnte für Indikatoren mit festem Referenzbereich der Referenzwert R das Shrinkage-Target sein. Geht man davon aus, dass es einen starken Zusammenhang zwischen

Fallzahl und Ergebnisqualität gibt, so können Shrinkage-Targets auch fallzahlabhängig gewählt werden (Ash et al. 2012, George et al. 2017).

Im Abschnitt 5.3.2 wurde bereits der Empirical-Bayes-Ansatz vorgestellt, bei dem die A-priori-Verteilung aus Daten mehrerer Leistungserbringer geschätzt wird. Da bei vielen QIs der Großteil der Leistungserbringerergebnisse im Referenzbereich liegt, führt dieser Ansatz tendenziell zu einem sehr wohlwollenden Shrinkage-Target. Da bei den meisten Leistungserbringern die Behandlungsqualität in Ordnung scheint, findet hier also die Qualitätssicherung unter der Grundannahme statt, dass auch beim nächsten betrachteten Leistungserbringer die Qualität vermutlich in Ordnung sein. Eine solche Annahme ist jedoch (anders als beim Public Reporting) in der Qualitätssicherung nicht unbedingt gerechtfertigt, wenn man sensitiv auf mögliche Qualitätsmängel reagieren will. Eine A-priori-Verteilung, bei der das Shrinkage-Target außerhalb des Referenzbereichs liegt, wie beispielsweise der Jeffreys-Prior, drückt dahingegen eine skeptische Grundhaltung aus. So eine A-priori-Verteilung basiert nicht auf empirischem Vorwissen, sondern auf Annahmen, mit denen man die quantitative Analyse der QS-Daten beginnen möchte. Daher lässt sich aus der Verwendung von Jefferys-Prior, welche einen A-priori-Erwartungswert für den Kompetenzparameter von ½ annimmt und der Nullhypothese $H_0: \theta \leq R$ eine A-priori-Wahrscheinlichkeit von $F_{\text{Beta}}(R|½, ½)$ zuordnet[45] nicht die Behauptung ableiten, dass bei der Mehrheit der Leistungserbringer ein Qualitätsmangel vorliegt.

5.3.3 Sensitivität und Spezifität der Einstufungsmethoden

Die Entscheidung, ob ein Leistungserbringerergebnis als quantitativ auffällig oder unauffällig eingestuft werden soll, ist immer eine Abwägung von Sensitivität und Spezifität bei der Klassifikation von θ: je strenger die Schwellenwerte einer Einstufungsmethode sind, desto höher ist die Sensitivität und desto niedriger die Spezifität, und vice versa. Für einen detaillierteren modellbasierten Vergleich von Sensitivität und Spezifität der Methoden werden im Folgenden vier Szenarien angenommen, bei denen ein Leistungserbringer jeweils die zugrunde liegende Rate von $\theta = 5\ \%$, $\theta = 10\ \%$, $\theta = 15\ \%$, und $\theta = 20\ \%$ hat, wiederum für einen Qualitätsindikator mit Referenzbereich $[0{,}10\ \%]$. Für diese vier Szenarien wird die Wahrscheinlichkeit einer quantitativ auffälligen Einstufung für die oben diskutierten Einstufungsmethoden berechnet, jeweils für Fallzahlen zwischen $J = 1, \ldots, 1000$. Die oberen beiden Darstellungen in Abbildung 17 zeigen dabei die Wahrscheinlichkeit falsch-positiver Einstufungen (1 - Spezifität) der Methoden, da die wahren zugrunde liegenden Raten innerhalb des Referenzbereichs liegen. Die unteren beiden Abbildungen zeigen die Wahrscheinlichkeit richtig-positiver Einstufungen (Sensitivität). Verglichen werden rechnerische Auffälligkeit, statistisch signifikante Auffälligkeit mit $\alpha = 5\ \%$ sowie die statistisch relevante Auffälligkeit mit $\zeta = 2$. Eine interaktive Version der Abbildung, bei der sowohl α als auch ζ variiert werden können, ist als Shiny-App unter https://iqtig.shinyapps.io/sensitivitaet_spezifitaet/ zu finden.

[45] Ist der Referenzwert z. B. $R = 10\ \%$, dann ist die A-priori-Wahrscheinlichkeit für die Nullhypothese z. B. 20,5 %.

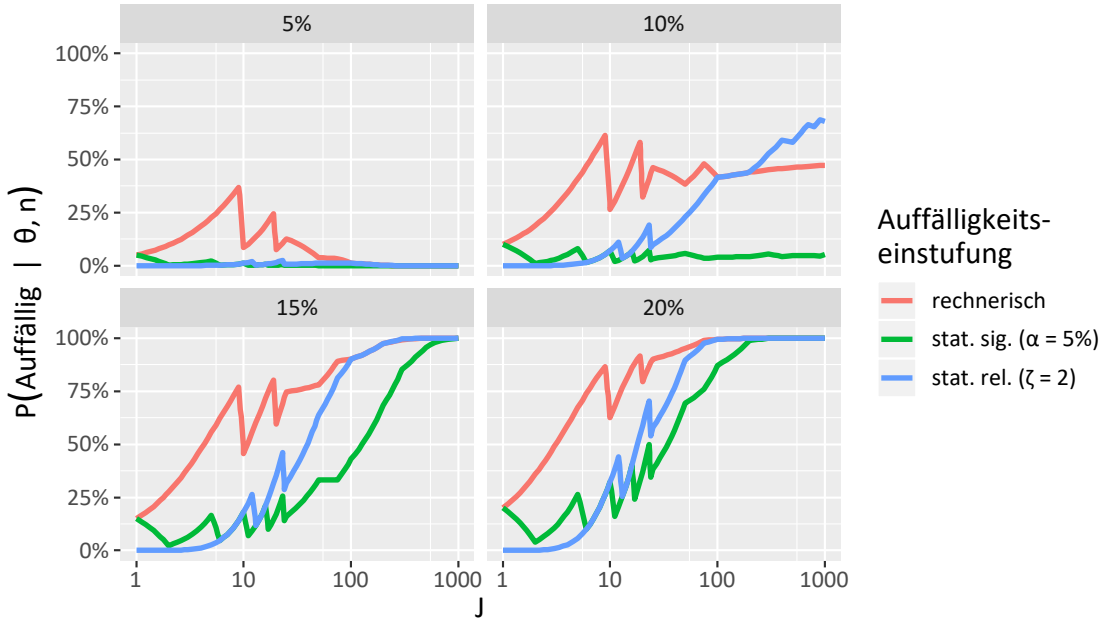

Abbildung 17: Vergleich von Sensitivität und 1-Spezifität für die drei Methoden der quantitativen Auffälligkeitseinstufung. Die Fallzahl auf der x-Achse ist dabei auf einer logarithmischen Skala dargestellt.

Der Fall, in dem die wahre zugrunde liegende Rate genau auf dem Referenzwert liegt, d. h. $\theta = 10\ \%$, ist insofern besonders, als die Methodik der statistisch signifikanten Auffälligkeit so ausgelegt ist, dass in diesem Fall (im Mittel über alle Fallzahlen J) die Wahrscheinlichkeit einer auffälligen Einstufung genau α entspricht, in diesem Fall $5\ \%$. Die Spezifität ist dabei wie auch im Fall $\theta = 5\ \%$ im Mittel über alle Fallzahlen bei der statistisch signifikanten Auffälligkeit sehr groß. Bei der statistisch relevanten Auffälligkeit dagegen ist die Spezifität bei $\theta = 5\%$ ebenfalls groß, bei $\theta = 10\ \%$ jedoch besonders bei großen Fallzahlen sehr klein und konvergiert mit steigender Fallzahl gegen 0. Umgekehrt hat die statistisch signifikante Auffälligkeit bei großen Fallzahlen die geringste Sensitivität der hier verglichenen Methoden. Der Grund für die schlechte Spezifität der statistischen Relevanz bei $\theta = 10\ \%$ für große Leistungserbringer ist, das große Leistungserbringer in diesem Grenzfall bei der statistischen relevanten Auffälligkeit schnell eine relevante Anzahl an Patienten und Patientinnen haben können, die über den Referenzwert hinaus ein unerwünschtes Qualitätsergebnis hatten. Hat ein großer Leistungserbringer jedoch einen Kompetenzparameter, der nur leicht unter R liegt, dann ist auch hier die Spezifität gut. Mit anderen Worten: Es spielt bei der statistischen Relevanz weniger eine Rolle, ob nun genau $\theta \leq R$ oder nicht, sondern der mit der Fallzahl gewichtete Abstand zum Referenzwert ist von Interesse.

Grundsätzlich bieten die statistisch signifikante und statistisch relevante Auffälligkeitseinstufung gegenüber der rechnerischen Auffälligkeit den Vorteil, dass der Trade-Off zwischen Sensitivität und Spezifität über einen Parameter (α, bzw. ζ) gesteuert und damit beispielsweise an QI-spezifische Besonderheiten angepasst werden kann, ohne den Referenzbereich verändern zu müssen. Diese Flexibilität erscheint vor dem Hintergrund der unterschiedlichen QS-Verfahren mit ihren unterschiedlichen Fallzahlen wünschenswert.

Die statistisch relevante Auffälligkeitseinstufung basierend auf ζ sowie die statistisch signifikante Auffälligkeit haben in den analysierten Szenarien beide die wünschenswerte Eigenschaft, dass ihre Spezifität bei wahrer zugrunde liegender Rate innerhalb des Referenzbereichs (aber nicht auf dem Referenzwert) mit steigender Fallzahl gegen 100 % und ihre Sensitivität bei zugrunde liegender Rate außerhalb des Referenzwerts gegen 100 % konvergieren. Wie bei der Betrachtung der Schwellenwerte im Funnelplot schon deutlich wurde, ist die statistisch signifikante Auffälligkeit mit $\alpha = 5\ \%$ dabei bei kleinen Fallzahlen strenger als die statistisch relevante mit $\zeta = 2$, bei großen Fallzahlen jedoch liberaler. Diese Abwägung zwischen strengeren Schwellenwerten für Leistungserbringer mit kleiner Fallzahl (statistisch signifikante Auffälligkeit) gegenüber strengeren Schwellenwerten für solche mit großer Fallzahl (statistisch relevante Auffälligkeit) leitet sich direkt aus den zugrunde liegenden Verlustfunktionen der beiden Methoden ab und ist zentral bei der Gegenüberstellung dieser beiden Klassifikationsmethoden.

5.3.4 Umgang mit verteilungsabhängigen Referenzwerten

Im Gegensatz zu fixen Referenzwerten definieren perzentilbasierte Referenzwerte nicht erwartbare Qualität (IQTIG 2019a). Sie erlauben lediglich einen Vergleich der Leistungserbringer anhand der beobachteten Leistungserbringerergebnisse. Für einen festen Referenzbereich R ist die interessierende Wahrscheinlichkeit $P(\theta_i > R|o_i, J_i, \dots)$ für den Leistungserbringer i. Wird anstelle eines fixen Referenzwertes lediglich ein Perzentil ξ vorgegeben, so interessiert die Wahrscheinlichkeit, dass der Kompetenzparameter des Leistungserbringers i zu den schlechtesten $\xi \times 100\ \%$ der Kompetenzparameter gehört, d. h.

$$P(\mathrm{Perc}(\theta_i) > \xi),$$

wobei

$$\mathrm{Perc}(\theta_i) = \frac{1}{I}\sum_{l=1}^{I} I\left(\theta_i \geq \theta_l\right)$$

der Perzentilrang[46] des Leistungserbringers i bzgl. seines Kompetenzparameters in der Population aller I Leistungserbringer ist. Dies kann auch als empirischen Quantil von θ_i in der Population der Werte $\theta_1, \dots, \theta_I$ aufgefasst werden. Wichtig ist auch die Erkenntnis, dass bei der analytischen Herangehensweise der Kompetenzparameter nicht direkt beobachtet wird, sondern nur anhand der beobachteten Daten im inferenz-statistischen Sinne geschätzt werden kann.

Während die Wahrscheinlichkeit $P(\theta_i > R|o_i, J_i, \dots)$ nur vom Leistungserbringer i und dem Referenzbereich R abhängt, fließen in die Wahrscheinlichkeit $P(\mathrm{Perc}(\theta_i) > \xi|o_1, J_1, \dots, o_I, J_I, \dots)$ auch die Ergebnisse anderer Leistungserbringer ein. Liegt beispielsweise in einem Indikator eine geringe Streuung der Ergebnisse vor, so ist die Unsicherheit bei der Bestimmung des Perzentil-Ranges größer als für Qualitätsindikatoren, die eine große Streuung der Indikatorergebnisse aufweisen. In der Praxis vieler gegenwärtiger Qualitätsindikatoren nach QSKH-RL wird häufig das 95. Perzentil als Referenzbereich gewählt. Unter Berücksichtigung von statistischer Unsicherheit

[46] Der Perzentilrang des Wertes 9 in der Menge {1,2, ... ,9,10} ist z. B. 2/10 = 20 %.

ist eine treffsichere Einstufung, ob der Leistungserbringer unter die schlechtesten 5 % der Leistungserbringer gehört, sehr schwer und setzt voraus, dass sich diese 5 % der Leistungserbringer stark von den restlichen 95 % absetzen, um diese sicher als „schlechteste 5 %“ zu klassifizieren. In den meisten Anwendungsfällen lässt sich aufgrund der beobachteten Daten nur in sehr wenigen Fällen sagen, dass ein Leistungserbringer mit hoher Sicherheit zu den schlechtesten 5 % gehört. Die Berücksichtigung von statistischer Unsicherheit mit festem Konfidenzniveau α bei gleichzeitiger Festlegung des 95. Perzentils als Referenzwert würde somit zu einem doppelten Toleranzmechanismus führen, der nur extrem ausreißende Leistungserbringerergebnisse als quantitativ auffällig klassifiziert.

Daher wird im Folgenden ein Vorgehen für die Bestimmung eines perzentilbasierten Referenzwerts vorgeschlagen, bei welchem ein vorher definierter Anteil an Leistungserbringerergebnissen quantitativ auffällig wird. Dies entspricht in etwa dem aktuellen Vorgehen bei der Bestimmung von verteilungsabhängigen Referenzwerten in der QSKH-RL, da auch dort prinzipiell auch ein bestimmter Anteil an Leistungserbringern quantitativ auffällig werden soll. Sei dafür $\{1, \dots, I\}$ die Menge der Leistungserbringer mit zugrunde liegendem Kompetenzparameter $\{\theta_1, \dots, \theta_I\}$ und q der Anteil an Leistungserbringerergebnissen, welche in einem Indikator quantitativ auffällig werden sollen. Da es sich immer um eine diskrete Menge an Leistungserbringern handelt, kann es vorkommen, dass der Anteil an quantitativen Auffälligkeiten nicht genau q sein kann, weshalb im Folgenden davon ausgegangen wird, dass *mindestens* $q \cdot I$ Leistungserbringerergebnisse quantitativ auffällig werden sollen. Weiterhin wird davon ausgegangen, dass es sich um einen Qualitätsindikator handelt, bei dem ein niedriges θ gute zugrunde liegende Qualität bedeutet.

Für die rechnerische Auffälligkeit wurde im Rahmen von (analytischen) Auswertungen[47] der Referenzwert nach der RAW20-Methode (Paddock 2014) bestimmt, indem das $q \cdot 100$. Perzentil der beobachteten Ergebnisse unter jenen Leistungserbringern ermittelt wird, die eine Grundgesamtheit von mindestens 20 Fällen haben:

$$R = \mathrm{Q}\left(q \cdot 100, \left\{\frac{o_i}{J_i}\right\}_{i \in H}\right),$$

mit Q als empirischer Quantilsfunktion und H als Menge aller Leistungserbringer mit $J_i \geq 20$. Die Grenze von 20 Fällen wird verwendet, damit die verwendeten beobachteten Leistungserbringerergebnisse halbwegs präzise Schätzungen der zugrunde liegenden Kompetenzparameter der jeweiligen Leistungserbringer sind. Diese Vorgehensweise kann jedoch, wie in Abschnitt 2.6 gezeigt, zu größeren Abweichungen zwischen der nominell festgelegten Anzahl an Auffälligkeiten $\lceil q \cdot I \rceil$ und der tatsächlichen Anzahl an Auffälligkeiten führen. Für enumerative Auswertungen ist die 20’er Fallzahl-Grenze nicht notwendig.

Für die statistisch signifikante Auffälligkeit im Rahmen von analytischen Auswertungen schlägt das IQTIG vor, dass der Referenzwert so gewählt wird, dass die A-posteriori-Wahrscheinlichkeit für die Nullhypothese für mindestens $q \cdot I$ Leistungserbringer kleiner gleich dem gewählten Schwellenwert α ist:

[47] Zum Beispiel Auswertungen nach QSKH-RL.

$$\min\left\{R : \sum_{i=1}^{I}[P(\theta_i \leq R \mid o_i, J_i) \leq \alpha] \geq q \cdot I\right\}.$$

Analog ist dazu für die statistische relevante Auffälligkeitseinstufung im Rahmen von Ratenindikatoren:

$$\min\left\{R : \sum_{i=1}^{I}\left[\int_{R}^{1} fn \cdot (\theta_i - R) \cdot J_i \cdot f_{\text{Beta}}(\theta \mid \ldots)d\theta \geq \zeta\right] \geq q \cdot I\right\}.$$

In beiden Fällen kann das Minimum über eine numerische Lösung des Optimierungsproblems bestimmt werden. Im Fall der statistisch signifikanten Auffälligkeit ist dies analog zur Bestimmung des $q \cdot 100$. Perzentils der unteren Grenze der zweiseitigen (1-2α) × 100 % Unsicherheitsintervalle aller I Leistungserbringer, d. h. zum Beispiel im Rahmen von Ratenindikatoren

$$R = \mathrm{Q}\big(q \cdot 100, \mathrm{F}_{\text{Beta}}(\alpha; \ldots)\big),$$

mit F_{Beta} als Verteilungsfunktion der A-posteriori-Wahrscheinlichkeit.

Gerade bei Qualitätsindikatoren, bei denen die beobachteten Ergebnisse der Leistungserbringer eine geringe Streuung aufweisen, kann es dabei bei der Bestimmung dieser Referenzwerte vorkommen, dass der verteilungsabhängige Referenzwert nah am oder sogar unterhalb des Bundesdurchschnitts liegt. Dies ist bei der statistisch signifikanten und statistischen relevanten Auffälligkeit eher der Fall als bei der rechnerischen Auffälligkeit, da bei letzterer keine Stochastizität der beobachteten Daten (zugunsten der Leistungserbringer) berücksichtigt wird.

5.3.5 Vor- und Nachteile der Klassifikationsmethoden

Die rechnerische Auffälligkeit entspricht in enumerativen Studiendesigns ohne Unsicherheit bei der Bestimmung von θ der optimalen Entscheidungsstrategie. Die Annahme, dass der zugrunde liegende Parameter eines Leistungserbringers ohne Unsicherheit über die beobachteten Daten bestimmt werden kann, trifft jedoch im Regelfall nicht zu. Da die rechnerische Auffälligkeit diese Unsicherheit nicht berücksichtigt, führt sie – im Vergleich zur statistisch signifikanten bzw. statistisch relevanten Einstufung – zu verhältnismäßig vielen falsch-positiven Klassifikationen (unter der Annahme, dass α bzw. z in der Regel so gewählt werden, dass die Methoden der statistischen Auffälligkeit zu höheren Schwellenwerten führt, als die rechnerische Auffälligkeit). Dies wird auch durch die empirischen Ergebnisse der Hintergrundanalyse in Abschnitt 2.6 untermauert. Dieser niedrigen Spezifität der rechnerischen Auffälligkeit bzgl. der Klassifikation des Kompetenzparameters θ steht eine hohe Sensitivität gegenüber.

Die statistisch signifikante Auffälligkeit ist wiederum in analytischen Studiendesigns, bei denen nur Aufwände für falsch-positive und falsch-negative Klassifikationen angenommen werden, die optimale Entscheidungsstrategie. Sie ist außerdem ein in der QS etabliertes Klassifikationsverfahren (Spiegelhalter et al. 2012) und die frequentistische Version findet bereits in den Verfahren nach plan. QI-RL Anwendung. Darüber hinaus lassen sich Sensitivität und Spezifität über den Parameter α steuern und ggf. pro Qualitätsindikator oder QS-Verfahren individuell festsetzen. Die Wahl dieses Parameters stellt jedoch gleichzeitig eine Herausforderung dar, da hier anders

als bei der rechnerischen Auffälligkeit – implizit oder explizit – die Aufwände für falsch-positive gegenüber falsch-negativen Klassifikationen abgewogen werden müssen.

Dies trifft auch auf die statistisch relevante Klassifikationseinstufung zu, allerdings wird dort der Aufwand eines Stellungnahmeverfahrens abgewogen mit dem Aufwand eines Falls mit unerwünschtem Ereignis (über die vom Referenzbereich hinaus tolerierte Anzahl an Fällen). Diese Aufwandsabwägung berücksichtig, dass die Aufwände einer falsch-negativen Klassifikation unter Umständen von der Fallzahl des zu klassifizierenden Leistungserbringers abhängen; sofern dies der Fall ist, ist die statistisch relevante Auffälligkeitsklassifikation die optimale Entscheidungsstrategie, da sie die erwarteten Aufwände unter diesen Annahmen minimiert. Ihre Anwendung führt jedoch auch dazu, dass im Vergleich zur rechnerischen oder statistisch signifikanten Auffälligkeitseinstufung die Sensitivität bzgl. der Klassifikation des Kompetenzparameters bei kleinen Leistungserbringern vergleichsweise geringer ist, da diese aufgrund ihrer Fallzahl von geringer Relevanz sind. Umgekehrt ist die Spezifität bei Leistungserbringern mit großer Fallzahl gering, da deren Relevanz stark ins Gewicht fällt. Demgegenüber stehen eine hohe Spezifität bei kleinen Fallzahlen und eine hohe Sensitivität bei großen Fallzahlen, was sich aus der zugrunde liegenden Verlustfunktion ableitet.

Ein kritischer Punkt bei sowohl der statistisch signifikanten als auch der statistisch relevanten Auffälligkeitseinstufung ist, dass es, je nach Wahl der A-priori-Verteilung und der Tuning-Parameter der Methoden, Fallzahlen geben kann, bei denen Leistungserbringer unabhängig vom Ergebnis nicht auffällig werden können. Als Beispiel: bei der statistisch relevanten Auffälligkeitseinstufung mit $\zeta = 2$ kann ein Leistungserbringer erst bei einer Fallzahl von 3 überhaupt auffällig werden (3/3). Die Frage ist, wie man mit den Leistungserbringern unter diese Fallzahl umgehen möchte: akzeptiert man, dass hier im Rahmen eines Erfassungsjahres nicht genügend Hinweise für ein Qualitätsdefizit gesammelt werden können? Muss man stattdessen die Datenbasis erweitern? Oder versucht man für diese Leistungserbringer über eine andere Vorgehensweise, z. B. anhand einer qualitativen Prüfung, ggf. nur für eine Stichprobe der Leistungserbringer, zu einer Bewertung zu kommen? Eine weitere Alternative ist, anstelle die Qualität bei den Leistungserbringern im Rahmen eines Zweifachstichproben-basierten Audit-Verfahren zu untersuchen (Tenenbein 1970, Raats und Moors 2003). Die genaue Wahl der Verfahrensweise hängt hier sehr von den Zielen der Qualitätsbewertung ab und wird weiter in Abschnitt 5.6 diskutiert.

5.4 Berücksichtigung der Daten mehrerer Erfassungsjahre

Die Güte der im Abschnitt 5.3 dargestellten Methoden zur quantitativen Einstufung von Leistungserbringern lässt sich verbessern, indem zusätzlich Informationen über die vergangenen Erfassungsjahre mit betrachtet werden. Speziell bei Leistungserbringern mit kleinen Fallzahlen ist die Hoffnung, dass eine solche Vergrößerung der Datengrundlage zu einer sichereren quantitativen Bewertung führt. Betrachtet wird als Beispiel ein Leistungserbringer mit QI-Ergebnissen für einen Ratenindikator mit Referenzbereich $[0 - 10\ \%]$ wie in Abbildung 18.

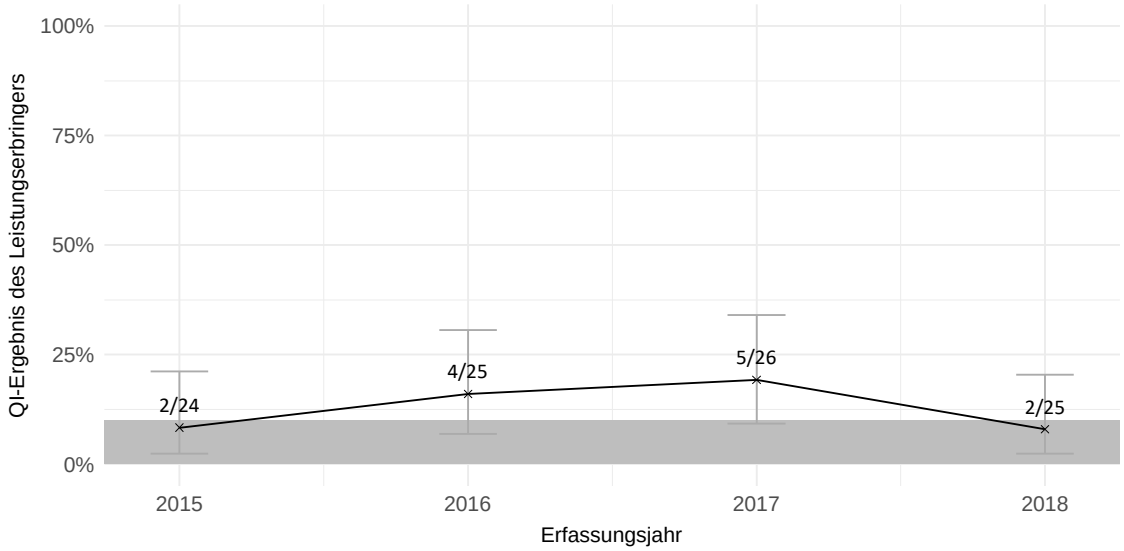

Abbildung 18: Beispielhafter Verlauf des QI-Ergebnis pro Erfassungsjahr eines hypothetischen Leistungserbringers für einen Ratenindikator.Der Referenzbereich ≤ 10 % ist grau schattiert. Um die Ergebnisse sind bayesianische 90 %-Unsicherheitsintervalle eingezeichnet.

Im Erfassungsjahr 2016 ist das QI-Ergebnis 4/25 = 16 % rechnerisch auffällig, jedoch ist das Ergebnis bei einem Schwellenwert von $\alpha = 0{,}05$ nicht statistisch signifikant auffällig (Wahrscheinlichkeit für die Nullhypothese nach der in Abschnitt 5.3.1.2 vorgestellten bayesianischen Methodik 0,1561). Im Erfassungsjahr 2017 ist das QI-Ergebnis 5/26 = 19 % erneut rechnerisch auffällig, jedoch ebenfalls nicht statistisch signifikant auffällig (Wahrscheinlichkeit für die Nullhypothese 0,0683). Im Erfassungsjahr 2017 ist der Leistungserbringer jedoch schon das zweite Mal in Folge rechnerisch auffällig. Außerdem ist das Gesamtergebnis 9/51 = 18 % für den 2-Jahres-Zeitraum 2016 bis 2017 statistisch auffällig (Wahrscheinlichkeit für die Nullhypothese 0,0429). Die Daten der 1-Jahres-Zeiträume liefern also jeweils wenig statistische Evidenz auf einen möglichen Qualitätsmangel, doch in den vereinigten Daten des 2-Jahres-Zeitraums ist die statistische Evidenz dafür größer, d. h. hier besteht ein Hinweis auf ein Qualitätsdefizit.

Bei der Bewertung ist zu berücksichtigen, dass die Daten aus größeren Erfassungszeiträumen nicht automatisch dazu führen, dass Qualitätsdefizite schneller entdeckt werden. Wird z. B. ein gleitendes Fenster von mehreren Erfassungsjahren für die Auffälligkeitseinstufung verwendet, würde eine große Verschlechterung im aktuellen Erfassungsjahr möglicherweise von guten Ergebnissen aus den früheren Jahren ausgeglichen. Gleichermaßen können alte, bereits behobene Qualitätsprobleme dazu führen, das aktuell gute Ergebnisse in den schlechten Ergebnissen der Vergangenheit untergehen. Eine Erweiterung der Datenbasis für die quantitative Auffälligkeitseinstufung um mehrere Erfassungsjahre bedingt daher auch eine gemeinsame Betrachtung der jährlichen Entscheidungen zur quantitativen Auffälligkeit und deren Auswirkungen.

5.4.1 Zusammenhang mit Strukturbruchproblemen, statistischer Prozesskontrolle und sequentiellen Entscheidungsproblemen

Es wird die folgende Notation verwendet: Die Zählvariable t läuft über die Erfassungsjahre. Die Anzahl an interessierenden Ereignissen im Erfassungsjahr t wird mit o_t bezeichnet, die Kardinalität der Grundgesamtheit mit J_t. Wie bereits im Abschnitt 5.2 wird die Modellannahme verwendet, dass o_t die Realisierung einer binomial $\mathrm{Bin}(J_t, \theta_t)$-verteilten Zufallsvariablen ist, wobei der Kompetenzparameter θ_t des Leistungserbringers ebenfalls vom Erfassungsjahr abhängen kann.

Das Entscheidungsproblem lässt sich formal als Strukturbruchproblem (*change point problem*) betrachten (Montgomery 2013, Tartakovsky et al. 2015). Verlässt θ_t den Referenzbereich, so wird dies als Strukturbruch interpretiert.[48] Der Strukturbruch soll möglichst zeitnah detektiert werden, weswegen sequentielle Methoden zur Anwendung kommen müssen (im Gegensatz zu retrospektiven Methoden). Anders als bei klassischen Strukturbruchproblemen sind nicht alle Parameteränderungen von Interesse, sondern vorwiegend solche, an denen sich θ_t von einem Wert im Referenzbereich zu einem Wert außerhalb des Referenzbereichs ändert.

Die Detektion solcher Strukturbrüche ist auch das Ziel der statistischen Prozesskontrolle (*statistical process control*, SPC) (Montgomery 2013). Bei gutartigen Parameterwerten (d. h. θ_t liegt im Referenzbereich) sagt man in der statistischen Prozesskontrolle auch, der Prozess sei „unter Kontrolle“, während bei ungünstigen Parameterwerten (d. h. θ_t liegt außerhalb des Referenzbereichs) der Prozess „außer Kontrolle“ ist.

Das Einflussdiagramm bietet wieder einen formalen Rahmen, um das sequentielle Entscheidungsproblem zu visualisieren und optimale Strategien zu identifizieren (Jensen und Nielsen 2007). Abbildung 19 zeigt die Erweiterung des in Abbildung 11 dargestellten Einflussdiagramms, in dem die jährliche quantitative Auffälligkeitseinstufung für einen analytisch auszuwertenden Ratenindikator als sequentielles Entscheidungsproblem dargestellt wird.

[48] Es kann natürlich sein, dass der Kompetenzparameter bereits beim ersten Zeitpunkt, an dem ein Leistungserbringer Daten für einen QI liefert, außerhalb des Referenzbereichs liegt. In diesem Fall wird formal dieser Startzeitpunkt des Verfahrens als Strukturbruch angesehen.

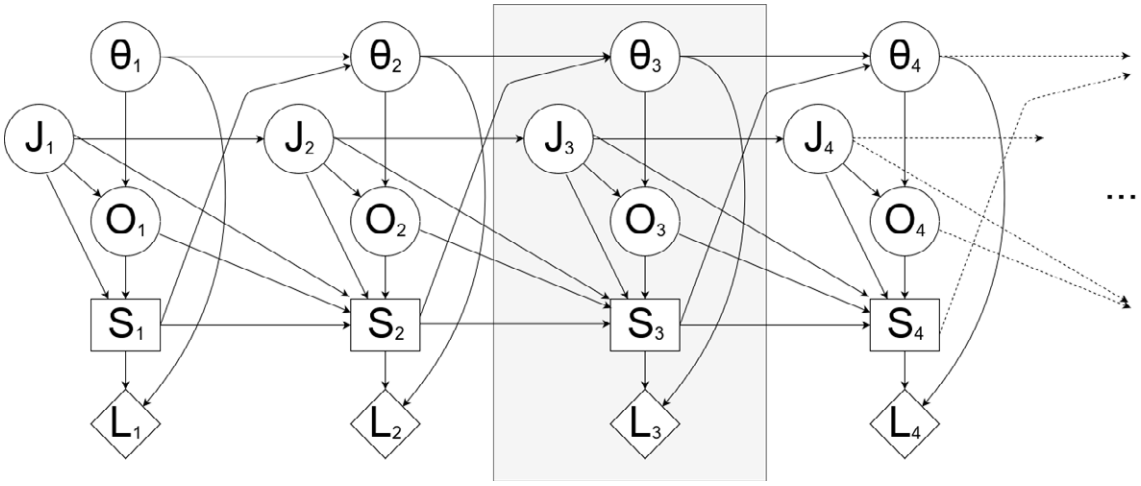

Abbildung 19: Entscheidungsdiagramm für das sequentielle Entscheidungsproblem der quantitativen Auffälligkeitseinstufung. Zeitpunkt 1 stellt das erste Erfassungsjahr mit Datenerfassung und Bewertung da. Die schattierte Fläche stellt die Knoten für ein ausgewähltes Erfassungsjahr ($t = 3$) dar.

In Abbildung 19 ist Zeitpunkt $t = 1$ das erste Jahr mit Datenerfassung und Bewertung für den Qualitätsindikator. Die quantitative Auffälligkeitseinstufung für dieses Jahr, S_1, wird anhand der beobachteten Anzahl an Fällen mit unerwünschtem Qualitätsergebnis, O_1 und der Gesamtanzahl an Fällen, J_1, getroffen. Die vorhandene beobachtete Evidenz (O_1, J_1) erlaubt somit Rückschlüsse auf θ_1, den Kompetenzparameter des Leistungserbringers in diesem Jahr. Ähnlich der Vorgehensweise in Abschnitt 5.2 wird basierend auf einer Abwägung verschiedener Aufwände eine quantitative Bewertung des Leistungserbringers vorgenommen, indem der Leistungserbringer für das Jahr als quantitativ auffällig ($S_1 = \text{ja}$) oder unauffällig ($S_1 = \text{nein}$) klassifiziert wird. Die unmittelbaren Aufwände dieser Klassifikation für das Erfassungsjahr sind in der Verlustfunktion L_1 festgehalten. Neu ist jedoch, dass die optimale Entscheidung nicht alleine auf der Abwägung der unmittelbaren Aufwände L_1 geschieht, sondern eine Betrachtung der Auswirkungen der Entscheidung und deren zukünftigen Aufwände notwendig sind. Dafür ist zum einen die zeitliche Dynamik des Kompetenzparameters θ_t des Leistungserbringers von Interesse, d. h. der Zeitreihe $\theta_1, \theta_2, \theta_3, \ldots$, und welche Entscheidungen in der Zukunft getroffen werden, d. h. nach welcher Strategie die Entscheidungen $S_2, S_3, \ldots$ getroffen werden. Um entsprechende optimale Entscheidungen treffen zu können ist es relevant zu wissen, wie stark sich der Kompetenzparameter von Jahr zu Jahr ändert; z. B. wie wahrscheinlich ist es, dass ein Leistungserbringer mit zureichender Qualität im nächsten Jahr sich verschlechtert bzw. ob es Leistungserbringer aus Eigeninitiative schaffen, Qualitätsdefizite zu beheben.

Wie aus dem Diagramm hervorgeht kann J_t prinzipiell über die Erfassungsjahre flukturieren. Zur Vereinfachung wird in den folgenden Betrachtungen jedoch J_t als fest und gegeben angenommen. Eine andere wichtige Komponente sind die Auswirkungen der quantitativen Auffälligkeitseinstufung (und der anschließenden fachlichen Bewertung) auf den Kompetenzparameter. Wie wahrscheinlich sind beispielsweise Verbesserungen nach einer quantitativen Auffälligkeit, wenn diese berechtigt ist? Alle diese Abhängigkeiten sind konzeptionell durch Pfeile im Diagramm illustriert. Eine konkrete Operationalisierung des Entscheidungsproblems durch das Dia-

gramm bedarf jedoch zusätzlich einer konkreten Festlegung der betrachteten Zustände der Variablen im Diagramm, den konkreten Übergangswahrscheinlichkeiten sowie den einzelnen Verlustfunktionen. Für ein Beispiel einer solchen Operationalisierung vgl. Abschnitt 5.4.4.

Prinzipiell liegen für jede Entscheidung im Diagramm immer alle Informationen aus der Vergangenheit vor, d. h. die Informationsgrundlage für S_1 besteht aus (O_1, J_1), für S_2 ist sie $(O_1, J_1, S_1, O_2, J_2)$, für S_3 ist die Grundlage $(O_1, J_1, S_1, O_2, J_2, S_2, O_3, J_3)$ etc. Daraus wird klar, dass die Strategien mit der Zeit immer komplexer werden können, weil die Strategie von immer mehr Variablen abhängt.

Zur Vereinfachung bei der Handhabung und somit gleichzeitig zur besseren Kommunizierbarkeit der Strategien wird im folgenden Text vereinfachend eine obere Grenze für den Einbezug der vergangenen Information verwendet. In der Abbildung 19 geht z. B. für die Entscheidung immer nur die Information aus dem aktuellen Jahr sowie die beobachtete Information aus dem Vorjahr ein, d. h. die Informationsgrundlage für S_3 ist nur $(O_2, J_2, S_2, O_3, J_3)$ – die frühere Vergangenheit wird vergessen. Dieses Vorgehen entspricht dem *Limited Memory Influence Diagram* (Lauritzen und Nilsson 2001) bzw. der Vorgehensweise bei *Partially Observed Markov Decision Processes* (Kaelbling et al. 1998, Åström 1965). Die Einschränkung auf zwei aufeinanderfolgender Erfassungsjahre soll nicht bedeuten, dass zu einem späteren Zeitpunkt (oder in der fachlichen Bewertung) nicht noch weitere Erfassungsjahre betrachten werden könnten. Je größer jedoch die Anzahl an betrachteten Jahren ist, desto größer ist die Rolle, die Trends in der medizinischen Versorgung, Veränderungen beim Leistungserbringer bzw. Veränderungen wegen früheren Qualitätsinterventionen spielen, sodass die Daten nur bedingt aussagekräftig für die aktuelle Kompetenz eines Leistungserbringers sind.

Für die Beschreibung von Entscheidungsstrategien, die auf den Daten zweier Erfassungsjahre beruhen, wird im Folgenden das zum Zeitpunkt der Entscheidung aktuelle Erfassungsjahr mit $t = 0$ bezeichnet, das Vorjahr entsprechend mit $t = -1$.

Eine grundlegende Idee für die Formulierung solcher 2-Jahres-Strategien ist nun die Kombination aus zwei Sachverhalten: Zum einen möchte man möglichst frühzeitig große Veränderungen im Kompetenzparameter detektieren können. Andererseits kann es kleinere Änderungen geben, die nur schwierig anhand der Daten eines Erfassungsjahres detektiert werden können. In dieser Situation ist es notwendig, die Daten o_0 und o_{-1} bzw. J_0 und J_{-1} gemeinsam zu betrachten. Die gepoolte Betrachtungsweise birgt jedoch die Gefahr, dass gute Ergebnisse im ersten Jahr schlechte Ergebnisse im zweiten Jahr ausgleichen können. Damit die gepoolte Betrachtungsweise der Daten tatsächlich zusätzliche Informationen liefert, muss allgemein davon ausgegangen werden, dass $\theta_{-1} \approx \theta_0$. Daher lohnt es sich, die Daten der beiden Erfassungsjahre sowohl separat als auch gemeinsam zu betrachten.[49] Vor allem aufgrund der zeitlichen Verzögerung, mit der die Ergebnisse der quantitativen Einstufung an die Leistungserbringer übermittelt und

[49] Eine ähnliche Situation tritt bei der Strukturbruchdetektion auf, wo große Parameteränderungen ebenfalls schneller zu detektieren sind. In der Strukturbruchdetektion werden ebenfalls gleitende Zeitfenster verwendet. Der Vorschlag, gleitende Zeitfenster unterschiedlicher Größe (1- und 2-Jahres-Zeiträume) parallel zu betrachten, ergibt sich daraus, dass man zumindest bei der Betrachtung des 1-Jahres-Zeitraums ohne die Hilfsannahme $\theta_{-1} \approx \theta_0$ auskommt und diese Hilfsannahme daher erst so spät wie möglich gemacht wird.

Stellungnahmeverfahren geführt werden (im Sommer des Folgejahres), ist zu vermuten, dass eine quantitative Auffälligkeit nur wenig Einfluss auf den zugrunde liegenden Kompetenzparameter im direkt anschließenden Erfassungsjahr hat. Daher wird davon abgesehen, das Ergebnis s_{-1} in der Strategie für s_0 zu verwenden.

Die Strategien werden nach folgendem Prinzip entworfen: Im ersten Schritt werden die QS-Daten des aktuellen Erfassungsjahres für den Leistungserbringer betrachtet. Liefern die Daten einen hinreichenden Hinweis auf ein mögliches Qualitätsdefizit des Leistungserbringers im aktuellen Erfassungsjahr, so wird der Leistungserbringer im ersten Schritt als „quantitativ auffällig" eingestuft. Liefern die Daten keinen hinreichenden Hinweis auf ein mögliches Qualitätsdefizit, so wird der Leistungserbringer als „quantitativ unauffällig" eingestuft. Legen die Daten hingegen den Verdacht nahe, dass ein Qualitätsdefizit bestehen könnte, ohne dass der Hinweis als ausreichend erachtet wird, um im Sinne der Effizienz des Verfahrens ein Stellungnahmeverfahren auszulösen (d. h. die Abweichung vom Referenzwert ist vermutlich gering), so werden die QS-Daten des Vorjahres hinzugezogen. Nun werden im zweiten Schritt für den Leistungserbringer die QS-Daten des Vorjahres und die QS-Daten des aktuellen Jahres gemeinsam betrachtet. Liefern die QS-Daten beider Jahre zusammen genug statistische Evidenz für ein mögliches Qualitätsdefizit, wird ein Stellungnahmeverfahren ausgelöst. Lassen die QS-Daten beider Jahre keinen Verdacht auf ein mögliches Qualitätsdefizit aufkommen, so wird der Leistungserbringer in Schritt 1 als „quantitativ unauffällig" eingestuft, d. h. es besteht kein Hinweis auf ein Qualitätsdefizit. Bei dieser Vorgehensweise spielen die QS-Daten des aktuellen Erfassungsjahres daher auch stets bei der Bewertung die wichtigste Rolle.

Insgesamt ist das Vorgehen motiviert durch Ideen aus der statistischen Prozesskontrolle sowie durch sequentielle Hypothesentests. Das zweistufige Vorgehen ähnelt einem zweifachen Prüfplan (vgl. Abschnitt 15.3.1 in Montgomery (2013)), wobei jedoch im Unterschied zu diesem keine Stichproben gezogen, sondern komplette Erfassungsjahre betrachtet werden.

Um das Verfahren zu konkretisieren, muss man festlegen, wie in jedem Schritt die betrachteten QS-Daten bewertet werden. Dazu sind zwei Festlegungen nötig:

- Unter welchen Bedingungen wird der Hinweis auf ein mögliches Qualitätsdefizit als ausreichend gesehen, um ein Stellungnahmeverfahren zu starten?
- Unter welchen Bedingungen liegt kein Verdacht auf ein Qualitätsdefizit vor?

Beispielsweise könnte man für den ersten Schritt des soeben beschriebenen Verfahrens, in dem lediglich die aktuellen QS-Daten betrachtet werden, folgende Festlegung treffen: Ist das QI-Ergebnis des aktuellen Erfassungsjahres statistisch signifikant auffällig beim vorgegebenen Signifikanzniveau, so gibt es statistische Evidenz für ein mögliches Qualitätsdefizit im aktuellen Erfassungsjahr. Ist das QI-Ergebnis dagegen rechnerisch unauffällig, so kann man definieren, dass wenig statistische Evidenz für ein mögliches Qualitätsdefizit im aktuellen Erfassungsjahr vorliegt.[50] Unabhängig von der Festlegung wird jedes Verfahren hier wegen stochastischer Einflüsse auch Fehler begehen: Leistungserbringer mit Qualitätsdefiziten könnten per Zufall trotzdem ein

[50] Der fehlende hinreichende Hinweis auf ein Qualitätsdefizit bedeutet natürlich keineswegs im Umkehrschluss, dass ein Hinweis für zureichende Qualität vorliegt.

gutes Ergebnis vorweisen und umgekehrt, vgl. Abschnitt 5.3.3. Die Frage ist, wie es zu einer guten Abwägung der verschiedenen Perspektiven kommen kann.

In den späteren Schritten muss man festlegen, wie man die QS-Daten aus mehreren Erfassungsjahren gemeinsam betrachtet. Dazu gibt es zwei prinzipielle Möglichkeiten:

- Man kann Bewertungen der QS-Daten der einzelnen Erfassungsjahre kombinieren. Beispielsweise könnte man im zweiten Schritt festlegen, dass ein hinreichender Hinweis auf ein Qualitätsdefizit vorliegt, falls das QI-Ergebnis des Leistungserbringers sowohl im aktuellen Erfassungsjahr als auch im Vorjahr rechnerisch auffällig ist.[51]
- Man kann die QS-Daten vereinigen und gemeinsam bewerten. Beispielsweise könnte man im zweiten Schritt festlegen, dass ein hinreichender Hinweis auf ein Qualitätsdefizit gegeben ist, wenn das über den gesamten 2-Jahres-Zeitraum berechnete QI-Ergebnis statistisch signifikant auffällig ist.

In beiden Fällen kann dabei auf verschiedene bekannte Bewertungsmethoden für QS-Daten zurückgegriffen werden, wie rechnerische Auffälligkeit, statistische signifikante Auffälligkeit oder die im Abschnitt 5.3.1.3 eingeführte Einstufung nach der statistischen Relevanz. Bei der statistisch signifikanten Einstufung ist in diesem Abschnitt stets die in Abschnitt 5.3.1.2 vorgestellte bayesianische Methodik gemeint. Somit können die Mehrjahreserweiterungen in diesem Abschnitt als Erweiterungen nach dem Baukastenprinzip gesehen werden, bei denen die Verfahren aus Abschnitt 5.3 kombiniert werden.

5.4.2 Untersuchte Methoden

In diesem Abschnitt werden einige konkrete Verfahren beschrieben. Ein Vergleich der Methoden findet in den Abschnitten 5.4.3 bis 5.4.4 statt.

5.4.2.1 Prüfung, ob kein Verdacht auf ein mögliches Qualitätsdefizit vorliegt

In allen Methoden wird die Einschätzung, ob anhand der vorliegenden Daten ein Verdacht auf ein Qualitätsdefizit im aktuellen Erfassungsjahr besteht, mithilfe der rechnerischen Auffälligkeitseinstufung getroffen. Alternativ könnte man auch hier stochastische Einflüsse berücksichtigen und beispielsweise analog zur Methodik der statistisch signifikanten Auffälligkeit überprüfen, ob es statistische Evidenz (zu einem festzulegenden Schwellenwert α) dafür gibt, dass es kein Qualitätsdefizit gibt, d. h. im frequentistischen Rahmen ein Test der Art $\theta \geq R$ vs. $\theta < R$ bzw. im bayesianischen Rahmen eine Entscheidung für die Nullhypothese $\theta \leq R$. Ähnliche Verfahren werden beim sequentiellen Testen betrachtet. Die Verwendung der rechnerischen Auffälligkeit entspricht in gewisser Weise einem sehr schwachen Signifikanzniveau bzw. einem großen Schwellenwert für die A-posteriori-Wahrscheinlichkeit für die Nullhypothese.

[51] Hier ist zu beachten, dass die rechnerische Auffälligkeit an sich ein recht schwaches Kriterium ist, welches für sich alleine genommen zu zu vielen Auffälligkeiten führt. Dadurch, dass man rechnerische Auffälligkeit in mehreren aufeinanderfolgenden Jahren verlangt, verschärft man das Kriterium; vgl. die Diskussion in den Abschnitten 5.4.2.2 und 5.4.3.

Üblicherweise nimmt man jedoch beim sequentiellen Testen an, dass der betrachtete Prozess stationär ist. Bei den betrachteten Leistungserbringern ist dies, wie bereits erwähnt, nicht unbedingt der Fall. Ein Leistungserbringer, der im aktuellen Erfassungsjahr nicht rechnerisch auffällig ist, bei dem dennoch ein Stellungnahmeverfahren gestartet wird, könnte stets als Erklärung vorbringen, dass es eine Qualitätsverbesserung gegeben habe und dass es keinen aktuellen Qualitätsmangel gebe. Außerdem sollte ein konsistentes Verfahren, welches im aktuellen Jahr trotz rechnerischer Unauffälligkeit ein Stellungnahmeverfahren auslöst, mit hoher Wahrscheinlichkeit bereits in einem der Vorjahre ein Stellungnahmeverfahren ausgelöst haben, sodass die interessierenden Fälle der Vergangenheit bereits in einem Stellungnahmeverfahren besprochen wurden (siehe auch Abschnitt 5.4.5.1).

Soll die Prüfung, ob kein Verdacht auf ein Qualitätsdefizit vorliegt, strenger gestaltet werden, so ist zu beachten, dass dadurch eine größere Datengrundlage benötigt wird, um zum Schluss zu kommen, dass kein Qualitätsdefizit vorliegt. Dies liegt daran, dass die Referenzbereiche in der Regel klein sind (d. h. der Referenzwert ist bei Ratenindikatoren nahe bei 0 bzw. 1). Dadurch ist es schwieriger, Evidenz dafür zu sammeln, dass θ im Referenzbereich liegt, als Evidenz dafür zu sammeln, dass θ außerhalb des Referenzbereichs liegt. Als Beispiel wird ein Referenzbereich von $[0{,}10\ \%]$ betrachtet. Bereits bei einer Grundgesamtheit von lediglich $J = 1$ Fall, welcher interessierend ist, liegt die Wahrscheinlichkeit für die $H_0: \theta \leq 10\ \%$ bei 0,01. Hingegen liegt die Wahrscheinlichkeit für die umgedrehte Nullhypothese $H_0: \theta \geq 10\ \%$ immer über 0,05, solange die Grundgesamtheit kleiner ist als 18. Während also ein einzelner interessierender Fall bereits einen Hinweis auf ein mögliches Qualitätsdefizit liefern kann, benötigt der Leistungserbringer mindestens 18 Fälle ohne interessierendes Ereignis, um zum gleichen Schwellenwert statistische Evidenz für das Einhalten des Referenzbereichs zu liefern.

5.4.2.2 Laufregel

Betrachte die folgende Entscheidungsregel, die die QS-Daten eines Leistungserbringers aus zwei Erfassungsjahren verwendet:

- Ist das QI-Ergebnis des Leistungserbringers im aktuellen Erfassungsjahr statistisch auffällig, so wird der Leistungserbringer als „quantitativ auffällig“ eingestuft.
- Sind die QI-Ergebnisse des Leistungserbringers sowohl im aktuellen Erfassungsjahr als auch im Vorjahr rechnerisch auffällig, so wird der Leistungserbringer als „quantitativ auffällig“ eingestuft.
- Andernfalls wird der Leistungserbringer als „quantitativ unauffällig“ eingestuft.

Diese Entscheidungsregel ist von der Form, wie sie in der Einleitung vorgestellt wurde. Sie wird im Folgenden *Laufregel* genannt, da ähnliche Regeln unter diesem Namen z. B. in der statistischen Prozesskontrolle Anwendung finden, vgl. Abschnitt 5.3.5 in Montgomery (2013).

Formal lässt sich die Laufregel darstellen als

$$s_{\text{lauf2}}(o_0, o_{-1}, J_0, J_{-1}, R, \alpha) = s_{\text{stat.sig.Bayes}}(o_0, J_0, R, \alpha) \vee (s_{\text{rech}}(o_0, J_0, R) \wedge s_{\text{rech}}(o_{-1}, J_{-1}, R)).$$

Im Beispiel aus Abbildung 18 wäre der Leistungserbringer also im Erfassungsjahr 2017 quantitativ auffällig, da im zweiten Jahr in Folge eine rechnerische Auffälligkeit besteht.

Die Abbildung 20 illustriert, welche Ergebnisse zur Auslösung führen. Dabei wird ein Leistungserbringer mit $J = 25$ Fällen im Indikator und ein Referenzbereich $\leq 10\ \%$ angenommen. Die Punkte in den Abbildungen entsprechen möglichen Ergebniskombinationen von Vorjahr und aktuellem Jahr. Sie sind nach der quantitativen Einstufung nach der Laufregel eingefärbt. Die durchgezogene Linie ist die Grenze zur rechnerischen Auffälligkeit, die gestrichelte Linie die Grenze zur statistischen Auffälligkeit bei dem derzeit verwendeten Signifikanzniveau von $\alpha = 5\ \%$. Da die Laufregel die QS-Daten jedes Jahres isoliert betrachtet, setzt sich die Entscheidungsgrenze zwischen „quantitativ unauffällig“ und „quantitativ auffällig“ aus stückweise horizontalen und vertikalen Linien zusammen.

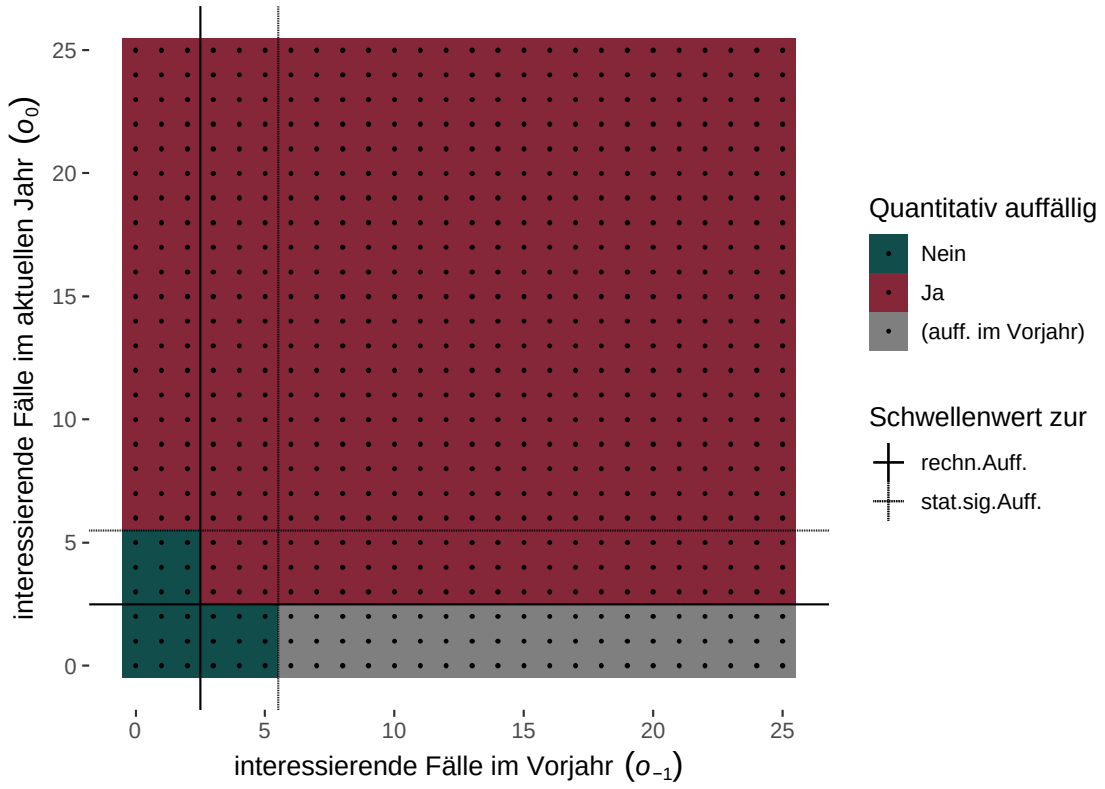

Abbildung 20: Verhalten der Laufregel bei verschiedenen Ergebnissen aus zwei Jahren für einen Leistungserbringer mit 25 Fällen in jedem Jahr

Beispielsweise wird ein Leistungserbringer, der im aktuellen Erfassungsjahr und im Vorjahr jeweils $J = 25$ Fälle in der Grundgesamtheit des QI hatte, als „quantitativ unauffällig“ klassifiziert, wenn es im Vorjahr zwei interessierende Ereignisse gab und im aktuellen Erfassungsjahr fünf (also insgesamt sieben interessierende Ereignisse in beiden Jahren). Gab es hingegen in beiden Jahren jeweils mindestens drei interessierende Ereignisse (d. h. insgesamt sechs), so wird der Leistungserbringer wegen der wiederholten rechnerischen Auffälligkeit als „quantitativ auffällig“ klassifiziert.

Bei den klassischen Laufregeln in der statistischen Prozesskontrolle (z. B. die Western Electric-Regeln, vgl. Abschnitt 5.3.5 in Montgomery (2013)) werden in der Regel Häufungen von Ereig-

nissen gezählt, die wesentlich seltener sind als rechnerische Auffälligkeiten bei Leistungserbringern mit Kompetenz im Referenzbereich. Wie später im Abschnitt 5.4.3 gezeigt, führt die Laufregel zu einer vergleichsweise hohen Auslösewahrscheinlichkeit für ein Stellungnahmeverfahren. Eine einfache Art, die Auslösewahrscheinlichkeit zu verringern, erhält man, wenn man die Bedingung, dass die QI-Ergebnisse des Leistungserbringers in den zwei aufeinanderfolgenden Jahren größer als der Referenzwert R sein sollen, durch die Bedingung, dass die QI-Ergebnisse des Leistungserbringers in den zwei aufeinanderfolgenden Jahren größer als ein verschärfter Referenzwert R' (mit $R' > R$) sein sollen, ersetzt. Diese Methode soll im Folgenden nicht weiter verfolgt werden, da die Laufregel nicht nur den Nachteil einer großen Auslösewahrscheinlichkeit besitzt, sondern daneben auch eine starke Fallzahlabhängigkeit aufweist, wie im Abschnitt 5.4.3 ausgeführt wird.

5.4.2.3 Statistische Auffälligkeit auf mehreren Jahren

Eine weitere einfache Regel zur quantitativen Auffälligkeit verwendet zur Beurteilung der Evidenzlage bzgl. eines mögliches Qualitätsdefizits beim Leistungserbringer das Konzept der statistisch signifikanten Auffälligkeit (dabei ist hier wie auch im Rest dieses Abschnittes stets die Abschnitt 5.3.1.2 vorgestellte bayesianische Methodik gemeint) in den QS-Daten mehrerer Erfassungsjahre. Zur Beurteilung der statistischen Evidenz für einen möglichen Qualitätsmangel wird also die A-posteriori-Wahrscheinlichkeit dafür, dass der Kompetenzparameter im Referenzbereich liegt, verwendet. Für die Erweiterung auf Daten mehrerer Jahre soll entsprechend die A-posteriori-Wahrscheinlichkeit auf den vereinigen QS-Daten des Leistungserbringers berechnet werden. Als Kriterium, ob in den betrachteten QS-Daten überhaupt ein Verdacht auf ein mögliches Qualitätsdefizit besteht, wird die rechnerische Auffälligkeit, berechnet basierend auf der Vereinigung der betrachteten QS-Daten, verwendet.

Bei maximal zwei betrachteten Jahren lautet die Regel also:

- Ist das QI-Ergebnis des Leistungserbringers im aktuellen Erfassungsjahr statistisch signifikant auffällig, so wird der Leistungserbringer als „quantitativ auffällig“ eingestuft.
- Ist das QI-Ergebnis für die Daten im aktuellen Erfassungsjahr rechnerisch auffällig und über die Daten beider Erfassungsjahre zusammen statistisch auffällig, so wird der Leistungserbringer als „quantitativ auffällig“ eingestuft.
- Andernfalls wird der Leistungserbringer als „quantitativ unauffällig“ eingestuft.

Formal lässt sich diese Entscheidungsregel darstellen als

$$s_{\text{stat.sig.Bayes2}}(o_0, o_{-1}, J_0, J_{-1}, R, \alpha) =$$
$$(s_{\text{stat.sig.Bayes}}(o_0, J_0, R, \alpha) \vee s_{\text{stat.sig.Bayes}}(o_0 + o_{-1}, J_0 + J_{-1}, R, \alpha)) \wedge s_{\text{rech}}(o_0, J_0, R).$$

Im Beispiel aus Abbildung 18 wäre der Leistungserbringer also im Erfassungsjahr 2017 quantitativ auffällig, da im zweiten Jahr in Folge eine rechnerische Auffälligkeit besteht.

Die Abbildung 21 zeigt wieder für einen Ratenindikator mit Referenzbereich $\leq 10\,\%$, welche Ergebnisse für einen Leistungserbringer mit $J = 25$ Fälle pro Erfassungsjahr nach diesem Verfahren zur Auslösung von Stellungnahmen führen. Es sind die statistischen Einstufungen in den beiden Jahren sowie die statistische Einstufung über beide Jahre wichtig. Daher setzt sich die

Entscheidungsgrenze aus horizontalen und vertikalen Linien sowie aus einer Gegendiagonale ($o_0 + o_{-1} = \text{konst.}$) zusammen.

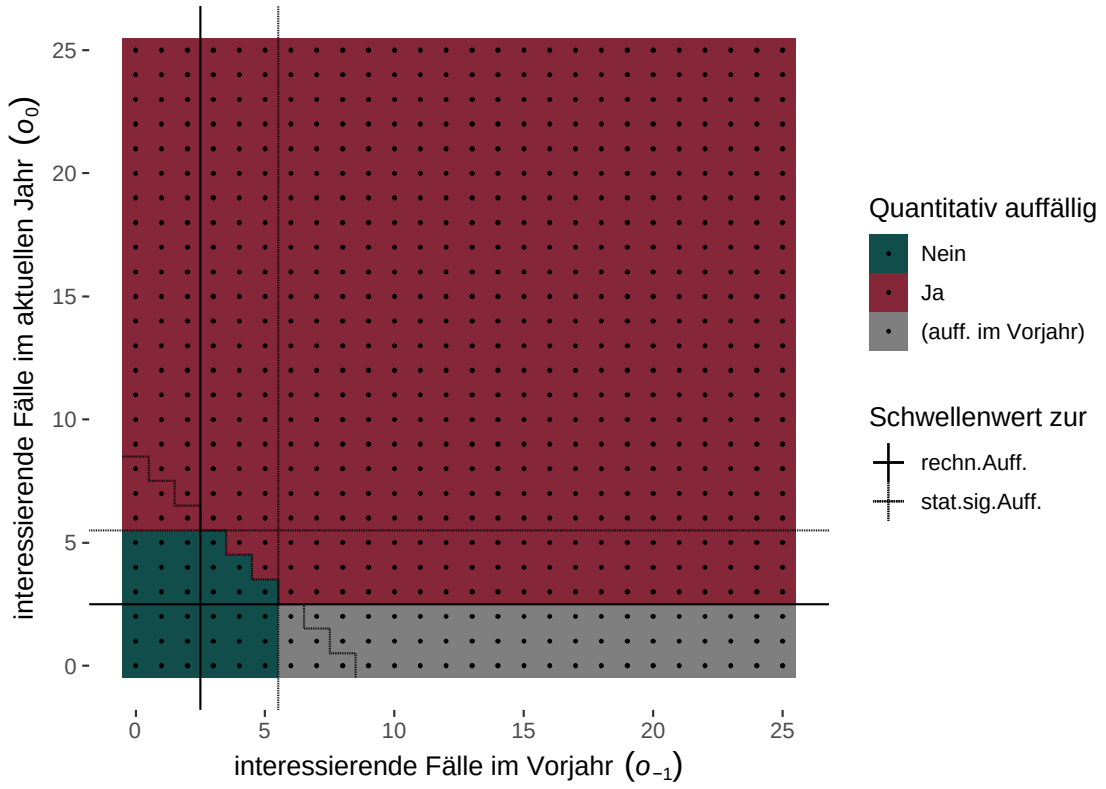

Abbildung 21: Verhalten der statistischen Auffälligkeit auf bis zu zwei Jahren bei verschiedenen Ergebnissen für einen Leistungserbringer mit 25 Fällen in jedem Jahr

Beispielsweise wird ein Leistungserbringer, der im aktuellen Erfassungsjahr und im Vorjahr jeweils $J = 25$ Fällen in der Grundgesamtheit des QI hatte, als „quantitativ unauffällig“ klassifiziert, wenn es im Vorjahr und im aktuellen Erfassungsjahr jeweils vier interessierende Ereignisse gab. Gab es hingegen im aktuellen Jahr vier interessierende Ereignisse und im Vorjahr fünf, so wird der Leistungserbringer als „quantitativ auffällig“ klassifiziert.

5.4.2.4 GLR-Binom-Verfahren

In der Statistischen Prozesskontrolle werden auch gerne CUSUM-Verfahren (Englisch: *cumulative sum*, Deutsch: kumulierte Summe) verwendet (Montgomery 2013, Abschnitt 9.1). In CUSUM-Verfahren wird für jeden einzelnen Zeitschritt eine Kennzahl berechnet, die ein Maß für die Evidenz ist, dass der Prozess „außer Kontrolle“ ist. Aus diesen Kennzahlen, die positiv oder negativ sein können, werden kumulative Summen gebildet. Dabei werden negative Zwischensummen auf 0 zurückgesetzt, um zu vermeiden, dass aus alten Daten zu viel „Guthaben“ angesammelt werden kann. Falls die kumulierte Summe einen vorher festgelegten Schwellenwert überschreitet, wird ein Signal gegeben.

Um solch ein CUSUM-Verfahren klassischer Weise einzustellen, muss man sich im Vorneherein auf zwei Hypothesen an die zugrunde liegende Rate θ an interessierenden Ereignissen festlegen:

Die Nullhypothese $H_0: \theta = \theta_{\mathrm{uk}}$ modelliert die Annahme, dass der Prozess „unter Kontrolle“ ist, während die Alternativhypothese $H_1: \theta = \theta_{\mathrm{ak}}$ (mit $\theta_{\mathrm{ak}} > \theta_{\mathrm{uk}}$) modelliert, dass der Prozess „außer Kontrolle“ ist. Dadurch ist implizit festgelegt, welche Größenordnung von Änderung im Parameter θ man detektieren will. Die Güte der Detektion hängt dabei stark von den getroffenen Festlegungen für θ_{uk} und θ_{ak} ab. Diese Formulierung der Nullhypothese und Alternativhypothese vereinfacht die Vorgehensweise der statistischen Auffälligkeit, bei der sowohl die Nullhypothese $\theta \leq R$ als auch die Alternativhypothese $\theta > R$ zusammengesetzer Natur ist, d. h. nicht einzelnen Werten, sondern Intervallen entspricht.

Im Folgenden soll stattdessen das GLR-Binom-Verfahren betrachtet werden, bei dem die erwartete Rate unter der Alternativhypothese nicht vorher festgesetzt werden muss, sondern für jeden Zeitpunkt aus den vorhandenen Daten geschätzt wird. Das GLR-Binom-Verfahren kann auch als eine sequentielle Aneinanderreihung von Likelihood-Quotiententests gesehen werden. In das oben eingeführte Verfahrensschema passt es wie folgt: Der Hinweis auf einen Qualitätsmangel in den QS-Daten eines Leistungserbringers in einem oder mehreren Erfassungsjahren wird als hinreichend erachtet, wenn ein Likelihood-Quotiententest die Nullhypothese $H_0: \theta \leq R$ zugunsten der Alternativhypothese $H_1: \theta > R$ ablehnt. Der Likelihood-Quotiententest hat als Parameter einen Schwellenwert $c > 0$. Die Nullhypothese wird abgelehnt, wenn der Logarithmus des Likelihood-Quotienten der beiden Hypothesen größer ist als c.

Insgesamt kann die Strategie auf folgende algorithmische Art dargestellt werden: Sei

$$s_{\mathrm{lrt}}(o, J, R, c) = I\left(\log \frac{f_{\mathrm{Bin}}\left(o; J, \frac{o}{J}\right)}{f_{\mathrm{Bin}}(o; J, R)} \geq c\right)$$

die Entscheidungsfunktion eines Likelihood-Quotiententests für die Nullhypothese $\theta \leq R$ mit Schwellenwert c. Damit ist

$$s_{\mathrm{glrbin2}}(o_0, o_{-1}, J_0, J_{-1}, R, c) =$$
$$(s_{\mathrm{lrt}}(o_0, J_0, R, c) \vee s_{\mathrm{lrt}}(o_0 + o_{-1}, J_0 + J_{-1}, R, c)) \wedge s_{\mathrm{rech}}(o_0, J_0, R),$$

vgl. z. B. Huang et al. (2012) bzw. Höhle und Paul (2008) .

Abbildung 22 zeigt wieder, welche Ergebnisse zur Auslösung führen (mit $J = 25$ und Referenzbereich $\leq 10\ \%$). Verwendet wird als Schwellenwert $c = 1{,}5$. Wie bei der statistischen Auffälligkeit auf zwei Jahren setzt sich die Entscheidungsgrenze aus horizontalen und vertikalen Linien sowie aus einer Gegendiagonale ($o_1 + o_2 = konst.$) zusammen.

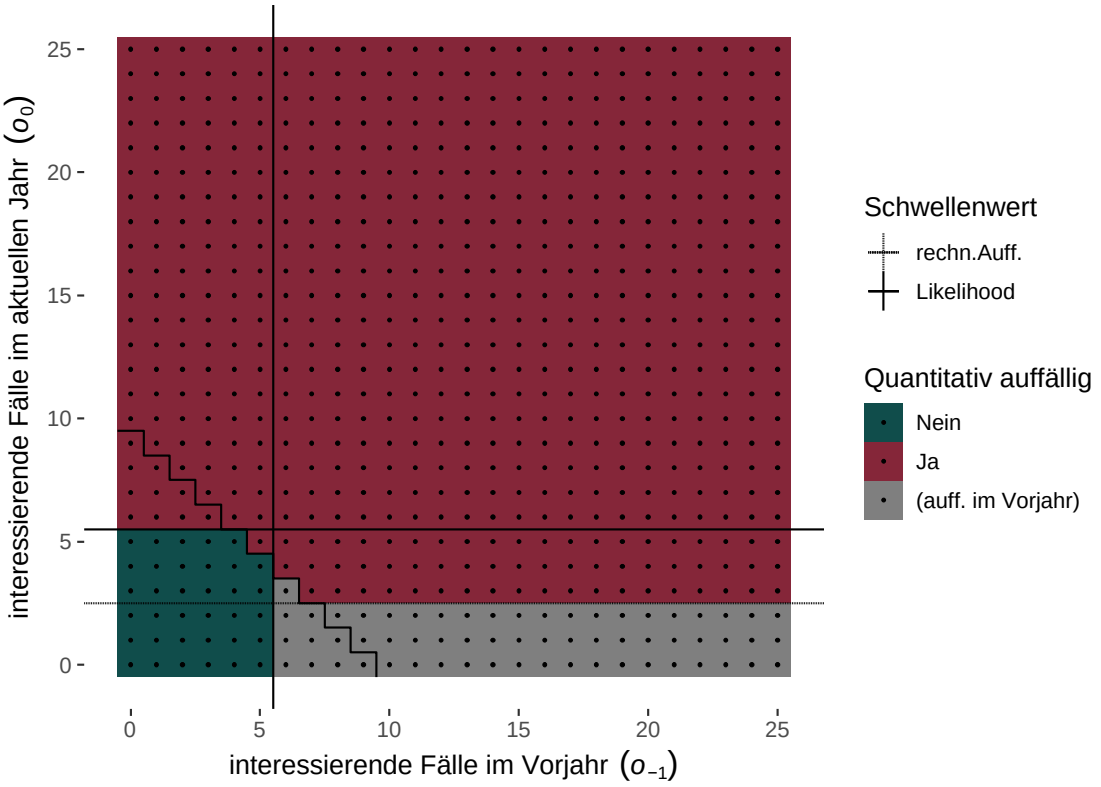

Abbildung 22: Verhalten des GLR-Binom-Verfahrens bei verschiedenen Ergebnissen aus zwei Jahren für einen Leistungserbringer mit 25 Fällen in jedem Jahr

Beispielsweise wird ein Leistungserbringer, der im aktuellen Erfassungsjahr und im Vorjahr jeweils $J = 25$ Fälle in der Grundgesamtheit des QI hatte, als „quantitativ unauffällig“ klassifiziert, wenn es im Vorjahr und im aktuellen Erfassungsjahr jeweils vier interessierende Ereignisse gab. Gab es hingegen im aktuellen Jahr und im Vorjahr jeweils fünf interessierende Ereignisse, so wird der Leistungserbringer als „quantitativ auffällig“ klassifiziert.

Vergleich von statistisch signifikanter Auffälligkeit und GLR-Binom-Verfahren

Das GLR-Binom-Verfahren ähnelt der statistisch signifikanten Auffälligkeit, die auf Basis mehrerer Erfassungsjahre berechnet wird: Während das GLR-Binom-Verfahren auf dem Likelihood-Quotienten-Test aufbaut, verwendet die frequentistische statistisch signifikante Auffälligkeit stattdessen einen bayesianischen Hypothesentest. In beiden Verfahren sind daher die Entscheidungsgrenzen ähnlich. Beide Verfahren besitzen einen Parameter (α bzw. c), über den sich Sensitivität bzw. Spezifität (pro Durchlauf des Verfahrens, also im esQS-Kontext pro Jahr) einstellen lassen. Bei der statistisch signifikanten Auffälligkeit besitzt der Parameter α eine einfache Interpretation als Signifikanzniveau. Diese Interpretation gilt jedoch nur für einen einzelnen Hypothesentest und beschreibt nicht das Gesamtverfahren, welches multiple Tests verwendet. Diese Interpretation kann auch ein Nachteil sein und unter Umständen eine freie Wahl des Parameters erschweren, zugunsten des aus plan. QI bekannten Parameterwerts $\alpha = 0{,}05$. Im vorliegenden Bericht wird daher der Parameter α auch nicht variiert.

5.4.2.5 Statistisch relevante Auffälligkeitseinstufung

Zur Beurteilung der statistischen Evidenz für einen möglichen Qualitätsmangel lässt sich auch die in Abschnitt 5.3.1.3 (Statistisch relevante Auffälligkeit) eingeführte Methodik verwenden, welche den Begriff „hinreichender Hinweis“ im Sinne der „Relevanz“ für Patienten und Patientinnen operationalisiert. Dazu wird aus den betrachteten QS-Daten der Erwartungswert $E[l(\cdot, y)]$ der entsprechenden Verlustfunktion geschätzt und anschließend mit einem vorgegebenen Schwellenwert ζ verglichen.

Bei der Berechnung von $E[l(\cdot, y)]$ wird angenommen, dass die Anzahl der Fälle, für die der Verlust geschätzt wird, gleich der Anzahl der Fälle im aktuellen Erfassungsjahr ist. Will man hingegen den erwarteten Verlust für ein Jahr abschätzen und verwendet man dafür die QS-Daten aus $k > 1$ Erfassungsjahren, so ist das Ergebnis $E[l(\cdot, y)]$ in diesem Fall durch k zu teilen. Alternativ kann man auch $E[l(\cdot, y)]$ mit $k \cdot \zeta$ vergleichen um festzustellen, ob der erwartete Verlust pro Erfassungsjahr größer gleich ζ ist.

Formal lautet die betrachtete 2-Jahres-Strategie also:

$$s_{\text{stat.rel2}}(o_{-1}, o_0, J_{-1}, J_0, \zeta, R) = (s_{\text{stat.rel}}(o_{-1} + o_0, J_{-1} + J_0, 2\zeta, R) \vee s_{\text{stat.rel}}(o_0, J_0, \zeta, R)) \wedge s_{\text{rech}}(o_0, J_0, R)$$

Die Abbildung 23 zeigt wieder, welche Ergebnisse zur Auslösung führen (mit $J = 25$ und Referenzbereich $\leq 10\ \%$). Verwendet wird als Schwellenwert $\zeta = 2$. Die Entscheidungsgrenze setzt sich aus einer horizontalen Linie sowie aus einer Gegendiagonale ($o_1 + o_2 = konst.$) zusammen.

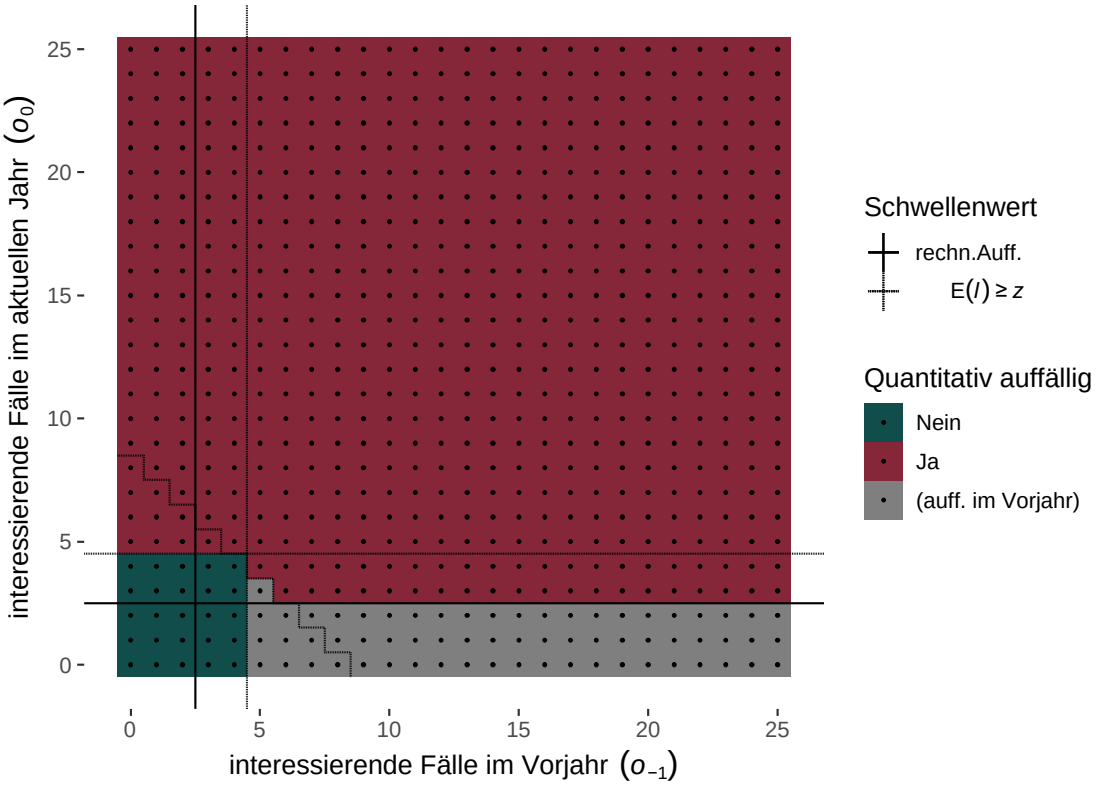

Abbildung 23: Verhalten der statistisch relevanten Auffälligkeitseinstufung auf den Daten zweier Jahre bei verschiedenen Ergebnissen für einen Leistungserbringer mit 25 Fällen in jedem Jahr

Beispielsweise wird ein Leistungserbringer, der im aktuellen Erfassungsjahr und im Vorjahr jeweils $J = 25$ Fälle in der Grundgesamtheit des QI hatte, als „quantitativ unauffällig“ klassifiziert, wenn es im Vorjahr und im aktuellen Erfassungsjahr jeweils vier interessierende Ereignisse gab. Gab es hingegen im aktuellen Jahr vier interessierende Ereignisse und im Vorjahr fünf oder mehr interessierende Ereignisse, so wird der Leistungserbringer als „quantitativ auffällig“ klassifiziert.

5.4.3 Vergleich anhand von Sensitivität und Spezifität

In diesem Abschnitt sollen die vorgestellten Methoden zur 2-Jahres-Einstufung mittels Sensitivität und Spezifität verglichen werden. Der nächste Abschnitt 5.4.4 kombiniert dann die beiden Dimensionen im Rahmen einer konkreten entscheidungstheoretischen Modellierung. Alternativ wäre es möglich, die Methoden an der Auslösewahrscheinlichkeit nach mehreren Jahren zu vergleichen. Man könnte z. B. fragen: Wie wahrscheinlich ist es, dass die Methode für einen gegebenen Beispielleistungserbringer innerhalb von 5 Jahren ein Stellungnahmeverfahren auslöst? Eine weitere Kennzahl, die in der statistischen Prozesskontrolle oft hinzugezogen wird, ist die erwartete Lauflänge (*average run length*, ARL), also die mittlere Anzahl von Jahren, nach denen ein Stellungnahmeverfahren ausgelöst wird. Hier wird lediglich die Auslösewahrscheinlichkeit pro Jahr betrachtet, da das Ziel der esQS die jährliche Bewertung von Leistungserbringern ist.

Die vorgeschlagenen Methoden mit ihren zugehörigen Tuning-Parametern unterscheiden sich in Sensitivität und Spezifizität (in egal welchem Zeitraum). Im Folgenden soll in Anlehnung an den Vergleich der 1-Jahres-Methoden in Abschnitt 5.3.3 die Wahrscheinlichkeit für das Auslösen

eines Stellungnahmeverfahrens (pro Durchlauf des Verfahrens, also pro Erfassungsjahr) für Beispielleistungserbringer anhand dreier Szenarien verglichen werden. Den Vergleich mittels Sensitivität und Spezifität pro Erfassungsjahr durchzuführen begründet sich darin, dass jährlich eine Entscheidung getroffen werden muss. Bei Beispielleistungserbringern, bei denen angenommen wird, dass der Kompetenzparameter im Referenzbereich liegt, d. h. $\theta \leq R$, entspricht diese Auslösewahrscheinlichkeit einer Fehlerwahrscheinlichkeit 1. Art und sollte möglichst klein sein. Bei Beispielleistungserbringern mit angenommener mangelhafter Kompetenz entspricht diese Auslösewahrscheinlichkeit der Power des Verfahrens und sollte möglichst groß sein. Die Parameter der Verfahren müssen so gewählt werden, dass ein guter Kompromiss zwischen geringer Fehlerwahrscheinlichkeit und großer Power erreicht wird.

Die Auslösewahrscheinlichkeit hängt neben der angenommenen Kompetenz des Leistungserbringers auch von der angenommenen Fallzahl des Leistungserbringers ab. Aufgrund der Diskretheit der Fallzahlen oszilliert die Auslösewahrscheinlichkeit mit der Fallzahl: Leistungserbringer mit gleicher Kompetenz und ähnlicher Fallzahl können sehr unterschiedliche Auslösewahrscheinlichkeit haben. Da dieser Effekt unerwünscht ist, ist ein weiteres Bewertungskriterium der verschiedenen Methoden, wie glatt die Abhängigkeit der Auslösewahrscheinlichkeit von der Fallzahl ist.

Im Folgenden wird lediglich der Parameter c der GLR-Binom-Methode variiert. Der Schwellenwert bei statistisch signifikanten Auffälligkeiten wird auf den auch im Verfahren plan. QI verwendeten Wert $\alpha = 0{,}05$ fixiert, und bei rechnerischen Bewertungen wird stets mit dem Referenzwert R verglichen. Der Parameter ζ der statistisch relevanten Auffälligkeiten wird auf den Wert $\zeta = 2$ fixiert. Über den Parameter c lassen sich die Fehlerwahrscheinlichkeiten 1. Art der GLR-Binom-Methode für einen Beispielleistungserbringer einstellen. Damit lässt sich diese Methoden mit den anderen Methoden anhand der Power vergleichen.

Als zweiter Vergleichsansatz, welche Sensitivität und Spezifität der Klassifizierung gegeneinander abwägt, wird in Abschnitt 5.4.4 eine zeitlich reduzierte Version des Entscheidungsdiagramms aus Abschnitt 5.4 mit nur zwei Zeitpunkten verwendet.

Betrachtet wird ein Ratenindikator mit Referenzbereich $\leq 10\ \%$. Wie in der Einleitung erwähnt, ist eines der Ziele der gemeinsamen Betrachtung mehrerer Erfassungsjahre die Verbesserung der Aussagekraft für kleine Leistungserbringer. Daher werden Fallzahlen bis maximal $J = 30$ betrachtet. Der Einfachheit halber wird angenommen, dass die Leistungserbringer in beiden aufeinanderfolgenden Jahren die gleiche Fallzahl aufweisen. Unterschiedliche Fallzahlen in beiden Jahren führen vermutlich zu etwas glatteren Kurven.

5.4.3.1 Abhängigkeit von der Fallzahl bei fester Kompetenz

Szenario: Kompetenz gleich Referenzwert

Für den Fehler erster Art wird zunächst beispielhaft der Fall eines Leistungserbringers mit $\theta = R = 10\ \%$ betrachtet. Es handelt sich also um einen Leistungserbringer mit gerade noch akzeptablem Kompetenzparameter. Die Auslösewahrscheinlichkeit ist hier also die Wahrscheinlich-

keit für den Fehler 1. Art im aktuellen Erfassungsjahr. Die Auslösewahrscheinlichkeit in Abhängigkeit der Fallzahl pro Jahr ist in Abbildung 24 dargestellt. Zusätzlich zu den in Abschnitt 5.4.2 eingeführten Methoden wird als Vergleich noch die statistisch signifikante Auffälligkeitseinstufung auf der Basis des aktuellen Erfassungsjahres sowie die statistisch relevante Auffälligkeitseinstufung auf der Basis des aktuellen Erfassungsjahres betrachtet.

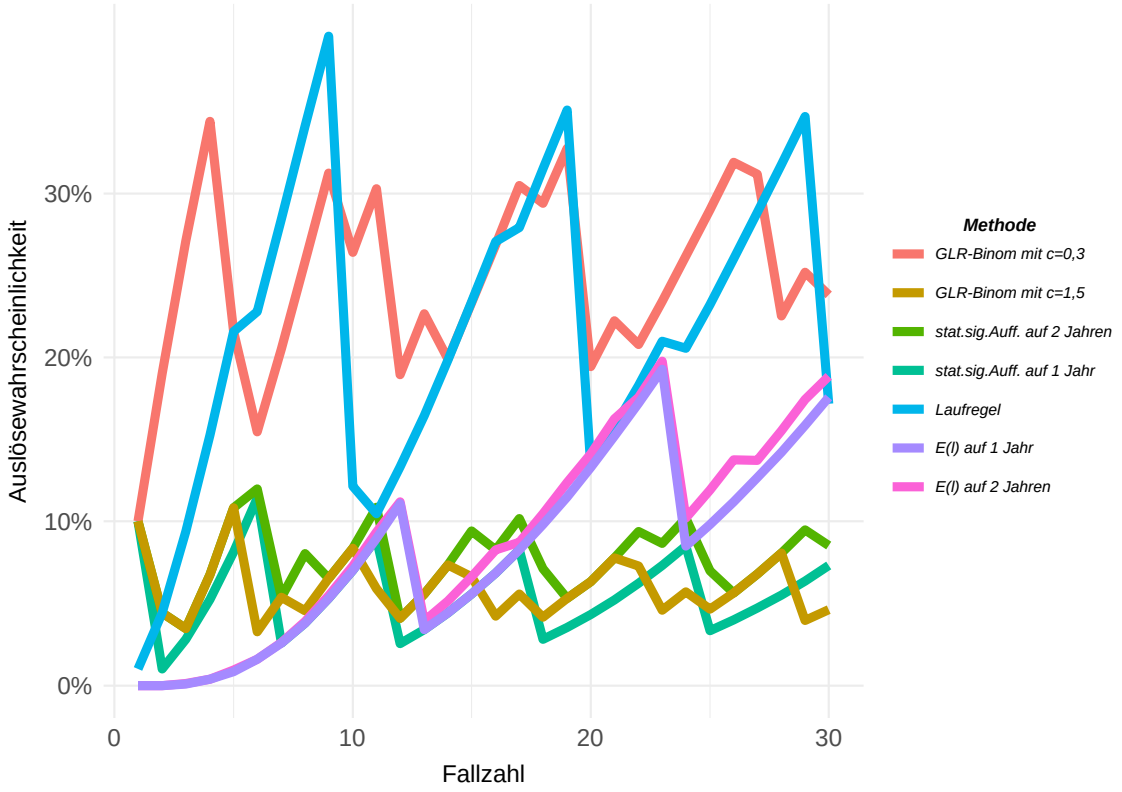

Abbildung 24: Die Auslösewahrscheinlichkeit verschiedener Methoden bei ϑ=R als Funktion von J.

Man sieht, dass die Wahrscheinlichkeit für einen Fehler 1. Art für die Laufregel am größten ist und außerdem die stärksten Oszillationen vorweist. Zum Teil hängt beides zusammen, d. h. die Oszillationen der Wahrscheinlichkeit sind größer, weil auch die Wahrscheinlichkeit absolut größer ist als die Wahrscheinlichkeit der anderen Methoden. Darüber hinaus kommen die großen Oszillationen auch durch die Form der Menge an Ereignissen, die zur Auslösung führen: Bei der quadratischen Form in Abbildung 20 können kleine Änderungen in der Fallzahl zu großen Änderungen in der berechneten Fehlerwahrscheinlichkeit führen. Zum Vergleich wird auch ein GLR-Binom-Verfahren mit recht niedrigem $c = 0{,}3$ betrachtet, welches eine vergleichbare durchschnittliche Fehlerwahrscheinlichkeit wie die Laufregel bietet, allerdings bei deutlich geringeren Oszillationen. Wie im Abschnitt 5.4.2.2 besprochen, lässt sich die Auslösewahrscheinlichkeit der Laufregel durch die Einführung eines weiteren Parameters $R' > R$ verringern. Dies würde jedoch nicht das Problem der Oszillationen lösen.

Die Fehlerwahrscheinlichkeit 1. Art der statistisch signifikanten Auffälligkeit über zwei Jahre, $s_{\text{stat.sig.Bayes2}}$, liegt leicht über der einfachen statistisch signifikanten Auffälligkeit. Beide Methoden haben ähnliche Eigenschaften wie die GLR-Binom-Methode mit $c = 1{,}5$. Tendentiell sind die Oszillationen für das GLR-Binom-Verfahren am geringsten.

Für die Einstufung nach der statistischen Relevanz steigt die Auslösewahrscheinlichkeit mit wachsender Fallzahl und ist für kleine Leistungserbringer sehr gering. Beide Effekte wurden bereits im Abschnitt 5.3.3 besprochen. Man sieht auch, dass es bei dieser Methodik vor allem bei kleinen Leistungserbringern keinen großen Unterschied gibt, ob man ein oder zwei Jahre betrachtet.

Szenario: $\theta = 2R$

Im Folgenden wird die Auslösewahrscheinlichkeit für Leistungserbringer mit unzureichendem Kompetenzparameter $\theta = 2R = 20\ \%$ geplottet (also die „Power“ für diesen Wert von θ):

Abbildung 25: Die Auslösewahrscheinlichkeit verschiedener Methoden bei ϑ=2R als Funktion von J.

Die Power der GLR-Binom-Methode mit $c = 1{,}5$ ist vergleichbar mit der Power der einfachen statistisch signifikanten Auffälligkeit und der statistisch signifikanten Auffälligkeit über zwei Jahre.

Szenario: $\theta = R/2$

Leistungserbringer mit $\theta = R$ haben „akzeptable“ Qualität; liegt der Bundesdurchschnitt im Referenzbereich, so bedeutet dies in der Regel, dass die meisten Leistungserbringer besser als der Referenzbereich sind. Daher wird als drittes ein Leistungserbringer mit $\theta = \frac{R}{2} = 0{,}05$ betrachtet.

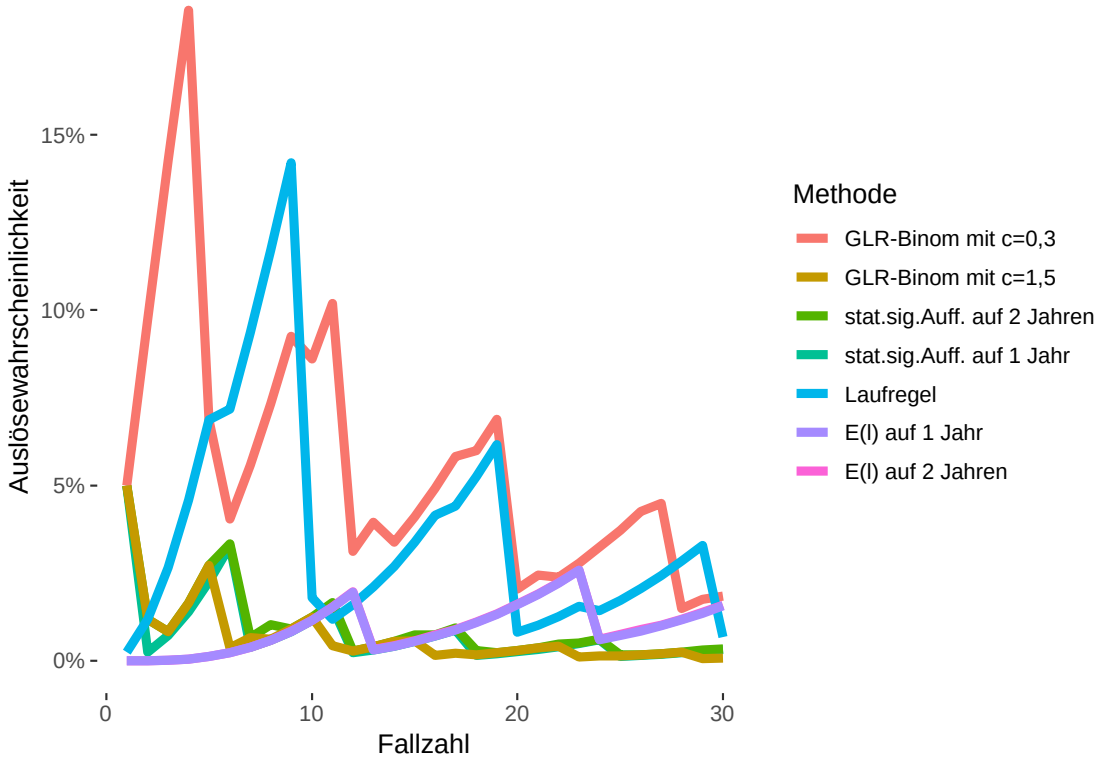

Abbildung 26: Die Auslösewahrscheinlichkeit verschiedener Methoden bei ϑ=R/2 als Funktion von J.

Bei den meisten Methoden geht die Auslösewahrscheinlichkeit mit steigender Fallzahl gegen Null. Die Laufregel und die Vergleichsmethode GLR-Binom mit $c = 0{,}3$ erreichen dabei jedoch relativ hohe Auslösewahrscheinlichkeiten (teilweise über 10 %). Die Auslösewahrscheinlichkeit der auf der statistischen Relevanz basierenden Methoden besitzen im betrachteten Fallzahlenbereich eine steigende Tendenz, bleiben jedoch klein (unter 3 %).

5.4.3.2 Abhängigkeit vom Kompetenzparameter bei fester Fallzahl

In diesem Abschnitt wird die Abhängigkeit der Auslösewahrscheinlichkeit vom Kompetenzparameter dargestellt, wobei die Fallzahl fixiert wird. Um die Oszillationen der Auslösewahrscheinlichkeit mit der Fallzahl etwas auszugleichen wird dabei über einen Bereich von Fallzahlen gemittelt. Wie gehabt wird ein Referenzwert von $10\ \%$ angenommen.

Abbildung 27 zeigt die Auslösewahrscheinlichkeit gemittelt für $J = 3, \ldots, 7$. Abbildung 28 zeigt die Auslösewahrscheinlichkeit gemittelt für $J = 11, \ldots, 20$.

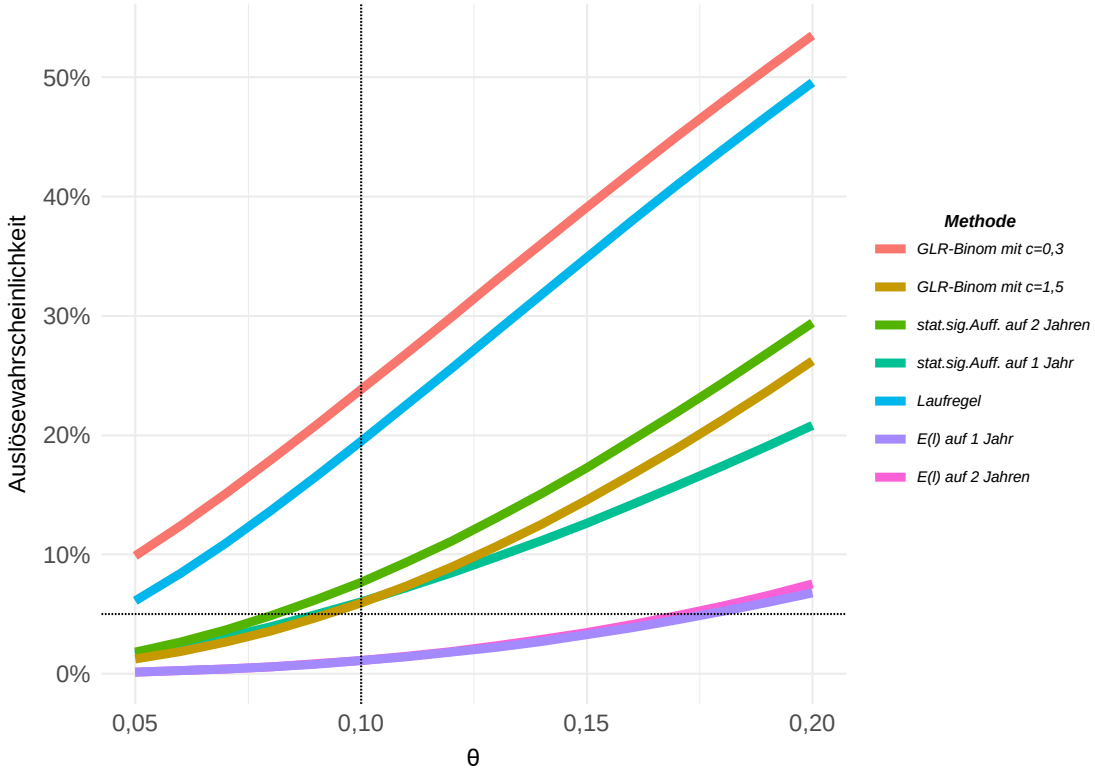

Abbildung 27: Die Auslösewahrscheinlichkeit verschiedener Methoden als Funktion von ϑ gemittelt über J=3,…,7. Die gestrichelten Linien markieren den Referenzwert R sowie den Schwellenwert α=0,05.

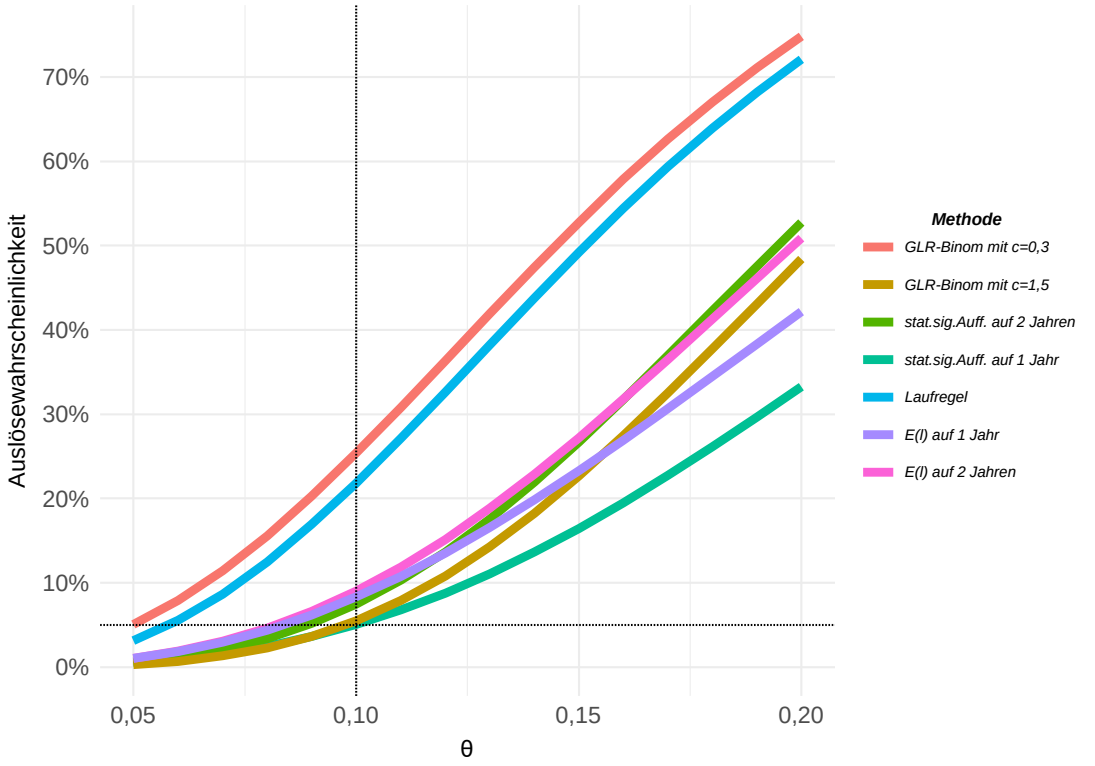

Abbildung 28: Die Auslösewahrscheinlichkeit verschiedener Methoden als Funktion von ϑ gemittelt über J=11,…,20. Die gestrichelten Linien markieren den Referenzwert R sowie den Schwellenwert α=0,05.

Man sieht, dass die Kurven bei den Verfahren, die nur ein einzelnes Erfassungsjahr berücksichtigen, jeweils langsamer ansteigen, als die Kurven der entsprechenden Verfahren, die zwei Erfassungsjahre berücksichtigen. Dies illustriert den Vorteil der Vergrößerung der Datengrundlage. Wie schon im vorangegangenen Abschnitt sichtbar ist die große Auslösewahrscheinlichkeit der Laufregel und der Vergleichsmethode GLR-Binom mit $c = 0{,}3$. Die Kurven für die statistisch signifikante Auffälligkeitseinstufung und das GLR-Binom-Verfahren mit $c = 1{,}5$ laufen parallel, was auf ähnliche Eigenschaften der Verfahren hindeutet.[52]

5.4.4 Vergleich durch Entscheidungsdiagramm mit zwei Zeitpunkten

Im vorherigen Abschnitt wurden jeweils Sensitivität und Spezifität der verschiedenen 2-Jahres-Auffälligkeitseinstufungsmethoden als separate Dimensionen betrachtet, was hilfreiche Einblicke liefert, aber es dem Entscheidungsträger überlässt, implizit zwischen diesen beiden Dimensionen zu gewichten. Für eine konkrete Empfehlung einer Methodik bedarf es, ähnlich der Vorgehensweise zur quantitativen Auffälligkeitseinstufung für die Daten eines Erfassungsjahres, der Kombination der Dimensionen im Rahmen einer entscheidungstheoretischen Modellierung.

Da generelle Betrachtungen zur optimalen Vorgehensweise im sequentiellen Kontext stark von den Annahmen zur z. B. zeitlichen Dynamik abhängen, sind generelle Aussagen zur optimalen Vorgehensweise deutlich schwieriger als für die 1-Jahres-Auffälligkeitseinstufung. Stattdessen werden die vorgeschlagenen Methoden im Rahmen eines konkreten Einflussdiagramms mit nur zwei Entscheidungen verglichen. In kalendarischer Zeitangabe sind die Entscheidungsvariablen im Einflussdiagramm unter dieser Annahme S_1 und S_2, die jeweils nach den vorgeschlagenen Strategien s_{lauf2}, $s_{\text{stat.sig.Bayes2}}$, $s_{\text{stat.rel2}}$, und s_{glrbin2} aus Abschnitt 5.4.2 getroffen werden sollen. Für die Entscheidung S_1 steht die Informationen bzgl. O_1 und J_1 zur Verfügung; für die Entscheidung S_2 stehen die Informationen bzgl. O_1, J_1, S_1, O_2 und J_2 zur Verfügung. Der Träger des Kompetenzparameters θ_t wird vereinfachend als die zwei möglichen Werte $\theta_{\text{unter-kontrolle}}$ bzw. $\theta_{\text{außer-kontrolle}}$ angenommen, wobei die Begriffe „unter-Kontrolle“ und „außer-Kontrolle“ für diese beiden Werte aus der statistischen Prozesskontrolle inspiriert sind (Montgomery 2013). Das heißt, das Ziel ist es, optimal für einen Leistungserbringer für jedes Erfassungsjahr differenzieren zu können, ob dieser die zugrunde liegenden Rate von $\theta_{\text{unter-kontrolle}}$ oder die zugrunde liegenden Rate $\theta_{\text{außer-kontrolle}}$ hat. Für die konkrete Berechnung wird $\theta_{\text{unter-kontrolle}}$ = 5 % und $\theta_{\text{außer-kontrolle}}$ = 20 %, gewählt. Als A-priori-Verteilung wird für θ_1 die folgende Verteilung gewählt,

$$P(\theta_1 = \text{unter-kontrolle}) = 90\ \%, \quad P(\theta_1 = \text{außer-kontrolle}) = 10\ \%.$$

Das heißt man geht davon aus, das 90 % der Leistungserbringer eine Kompetenz entsprechend dem Kompetenzparameter $\theta_{\text{unter-kontrolle}}$ haben und 10 % eine Kompetenzparameter von $\theta_{\text{außer-kontrolle}}$ haben.

[52] Für einen noch faireren Vergleich der Verfahren anhand des Verlaufs der Auslösewahrscheinlichkeiten müsste man die Parameter der Verfahren so wählen, dass die Auslösewahrscheinlichkeiten bei einem Vergleichswert $\theta = R$ nahezu identisch sind. Die Schlussfolgerungen bleiben jedoch gleich. Die entsprechenden Abbildungen werden hier nicht gezeigt, da, wie oben besprochen, in diesem Abschnitt lediglich der Parameter c variiert wird.

Für die Übergangswahrscheinlichkeiten $P(\theta_t \mid \theta_{t-1}, S_{t-1})$ des Kompetenzparameters werden folgende bedingte Wahrscheinlichkeiten gewählt:

Tabelle 10: Wahl der Übergangswahrscheinlichkeiten für den Kompetenzparameter

$\theta_t \setminus \theta_{t-1}, S_{t-1}$	Unter-Kontrolle		Außer-Kontrolle	
	S_{t-1}=nein	S_{t-1}=ja	S_{t-1}=nein	S_{t-1}=ja
Unter-Kontrolle	0,95	0,95	0	0,2
Außer-Kontrolle	0,95	0,95	1	0,8

Es wird also angenommen, dass ein Leistungserbringer, der im letzten Erfassungsjahr mit einem Kompetenzparameter entsprechend „unter-kontrolle“ gearbeitet hat, dies mit 95 % Wahrscheinlichkeit auch wieder im aktuellen Erfassungsjahr tut. Diese Wahrscheinlichkeit ist unabhängig davon, ob es im letzten Erfassungsjahr zu einer quantitativen Auffälligkeit kam oder nicht. War der Leistungserbringer dagegen im letzten Erfassungsjahr „außer-Kontrolle“, hängt die Situation im aktuellen Erfassungsjahr von der Entscheidung im letzten Erfassungsjahr ab: Hatte man sich damals gegen ein Signal entschieden (S_1 = nein), ist der Leistungserbringer weiterhin „außer-Kontrolle“. War die Entscheidung im letzten Jahr anstelle, (S_1 = ja), dann wird angenommen, dass mit einer Wahrscheinlichkeit von 20 % der Leistungserbringer im aktuellen Erfassungsjahr wieder „unter-Kontrolle“ ist.

Die Wahrscheinlichkeitsverteilung von $O_t|\theta_t, J_t$ entspricht wie immer einer Binomialverteilung, d. h.

$$P(O_t = o_t|\theta_t, J_t) = \binom{J_t}{o_t} \theta_t^{o_t}(1-\theta_t)^{J_t - o_t},$$

welches sowohl für $\theta_{\text{unter-kontrolle}}$ als auch $\theta_{\text{außer-kontrolle}}$ für den Träger von O_t , d. h. für 0, 1, …, J_t, berechnet werden muss. Die Verlustfunktionen sind in Analogie zu Abschnitt 5.3.1 jeweils zunächst entsprechend der Verlustfunktion beim statistisch-signifikanten Ansatz, welcher nur die Aufwände für Fehlklassifikationen betrachtet und in Tabelle 11 dargestellt ist.

Tabelle 11: Wahl der mit verschiedenen Fehlklassifikationen assoziierten Aufwände

Unter-Kontrolle	Außer-Kontrolle	Unter-Kontrolle	Außer-Kontrolle
S_{t-1}=nein	S_{t-1}=ja	S_{t-1}=nein	S_{t-1}=ja
0	fp	fn	0

mit $fp = 19$ und $fn = 1$.

Durch Lösen des Einflussdiagramms nach dem Algorithmus von Shenoy (1992) für einen Leistungserbringer mit $J_1 = J_2 = 25$ erhält man als optimale Strategie S_1, dass ein Signal generiert wird wenn $o_1 \geq 5$. Anhand des gleichen Lösungsalgorithmus findet man, dass das die optimale Entscheidung für S_2 ist, welche auch die Information zu S_1 (nein oder ja) mit einbezieht, die in Abbildung 29 dargestellte Form hat.

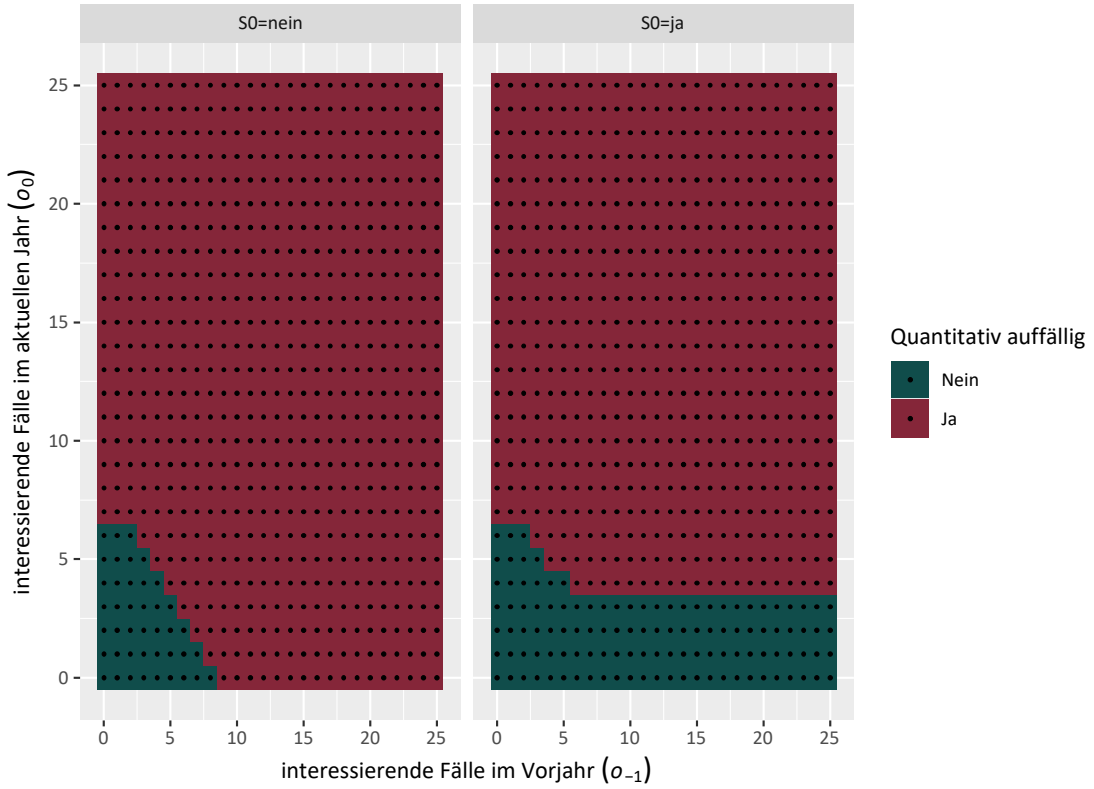

Abbildung 29: Optimale Strategie im gewählten Entscheidungsdiagramm in Abhängigkeit von o_0, o_{-1} und S_0. Links die Entscheidung in Abhängigkeit von o_0 und o_{-1}, falls $S_0 = nein$ und rechts die Entscheidung für den Fall, dass $S_0 = ja$.

Man sieht, dass die optimale Strategie für S_2 tatsächlich von S_1 abhängt. Hat man sich für eine quantitative Auffälligkeit bei S_1 entschieden und weil die Annahme ist, dass dies die zugrunde liegende Rate verbessert, würde man sich anschließend etwas zurückhaltender für eine quantitative Auffälligkeit bei S_2 entscheiden.

Innerhalb des gewählten Einflussdiagramms können die verschiedenen Strategien mittels des jeweiligen erwarteten Verlustes verglichen werden. Die oben erwähnte optimale Entscheidungsstrategie sowie die sehr vereinfachten Strategien immer bzw. niemals eine Auffälligkeit für sowohl S_1 als auch S_2 auszulösen, liefern entsprechende Baseline-Niveaus für den Vergleich. Die Ergebnisse finden sich in Tabelle 4.

Tabelle 12: Erwarteter Verlust der verschiedenen 2-Jahres-Methoden für das konkrete Einflussdiagramm mit zwei Zeitpunkten.

Methode	**erwarteter Verlust**
optimal	0,165
GLR-Binom (c=1,5)	0,183
stat.sig.bayes2 (α=0,005)	0,186
stat.sig.bayes2 (α =0,05)	0,186

Methode	erwarteter Verlust
nie auffällig	0,245
stat.rel2 (ζ=2)	0,350
lauf2	0,420
QSKH	4,304
immer auffällig	33,725

Per Konstruktion liefert die optimale Strategie den niedrigsten erwarteten Verlust. Aus der Tabelle wird klar, dass die Methode s_{stat2} mit α = 0,05 die besten Ergebnisse unter den in Abschnitt 5.4.2 vorgestellten Baukasten-Methoden für das betrachtete Setting liefert. Die Verwendung der rechnerischen Auffälligkeit in jedem Erfassungsjahr einzeln (die QSKH-RL-Methode) liefert wie erwartet schlechte Ergebnisse, weil es hier zu sehr vielen unnötigen Auffälligkeitseinstufungen kommt.

Insgesamt zeigen die obigen Ergebnisse, dass es prinzipiell möglich ist sehr komplexe Strategien zu verwenden, um im konkret angenommenen Einflussdiagramm optimale Entscheidungsstrategien zu erhalten. Jedoch sind diese stark abhängig von den verwendeten Annahmen, wie z. B. der A-priori-Verteilung für θ_1 und der Wahrscheinlichkeit, dass eine quantitative Auffälligkeit auch zu einer Qualitätsverbesserung führt. Auch das von der statistischen Prozesskontrolle inspirierte GLR-Binom-Verfahren liefert hervorragende Ergebnisse, jedoch ist das Verfahren bereits relativ komplex und es gibt keine gute Methodik zur Wahl des Schwellenwertes. Somit erscheinen die etwas einfacheren 2-Jahres-Einstufungen nach dem Baukasten-Prinzip anhand von statistischen Verfahren und ohne Berücksichtigung früherer Entscheidungen als robuste Alternativen, die auch gut in anderen Settings als dem betrachteten Einflussdiagramm funktionieren. Des Weiteren sind diese Strategien auch einfacher zu kommunizieren und liefern fast so gute Ergebnisse wie die optimale Strategie oder das GLR-Binom-Verfahren im konkreten Setting.

Bei der Interpretation der obigen Ergebnisse ist zu beachten, dass das konkrete Entscheidungsdiagramm für einen Leistungserbringer mit J = 25 Fällen und konkreten Festlegungen für die Wahrscheinlichkeitstabellen nur eine von vielen möglichen Einflussdiagrammen für einen Vergleich der Methoden ist. Trotzdem ist diese Art der Bewertung hilfreich, um zu einer kombinierten Sichtweise von Sensitivität und Spezifität zu kommen und liefert wichtige Einblicke in die Verhaltensweisen der verschiedenen Methoden.

5.4.5 Weitere Aspekte

5.4.5.1 Berücksichtigung vorangegangener Stellungnahmeverfahren

Bei der Verwendung von QS-Daten aus Vorjahren will man vermeiden, dass schlechte Ergebnisse in der Vergangenheit zu lange einen Einfluss auf die Gesamteinstufung haben. Einerseits ist dies dadurch gewährleistet, dass vor Erweiterung der Datengrundlage stets geprüft wird, ob die aktuell betrachteten Daten überhaupt einen Verdacht auf einen möglichen Qualitätsmangel nahelegen. Dadurch wird verhindert, dass ein Leistungserbringer mehrfach aus identischen Gründen

in ein Stellungnahmeverfahren gelangt. Daneben könnte man auch die Ergebnisse vorangegangener Stellungnahmeverfahren berücksichtigen.

Wenn im Vorjahr bei einem Leistungserbringer ein Stellungnahmeverfahren ausgelöst wurde, bei dem der Leistungserbringer bei der Bewertung dann abschließend als „unauffällig" eingestuft wurde, so könnte man dies als Anlass sehen, zur Bewertung des aktuellen Jahres keine Vorjahresdaten hinzuzunehmen. Da im Vorjahr ein Stellungnahmeverfahren ausgelöst wurde, enthalten die Vorjahresdaten scheinbare statistische Evidenz für einen möglichen Qualitätsmangel. Diese statistische Evidenz hat sich jedoch im Stellungnahmeverfahren nicht erhärtet. Daher sollte man diese statistische Evidenz nicht unbedingt zur Bewertung der aktuellen QI-Ergebnisse heranziehen. Dabei muss man jedoch vorsichtig die Gründe analysieren, die zu der Bewertung des Leistungserbringers geführt haben. Zum Beispiel könnte man bei der Bewertung, dass es sich bei den interessierenden Fällen im Vorjahr vermutlich um „tragische Einzelfälle" gehandelt haben könnte, durchaus die Evidenz für einen möglichen Qualitätsmangel in den QS-Daten des Vorjahres erneut berücksichtigen, da die Bewertung, ob es sich tatsächlich um Einzelfälle gehandelt hatte wiederum im Licht der neueren QS-Daten überprüft werden müsste.

Oft ist es wünschenswert, bei Leistungserbringern, bei denen im Vorjahr ein Stellungnahmeverfahren stattfand, ein genaues Augenmerk auf das Behandlungsergebnis zu haben, vor allem, wenn im Stellungnahmeverfahren ein Qualitätsdefizit festgestellt wurde. Bei Betrachtung der Vorjahresdaten ist dies in gewisser Weise gegeben: Das Stellungnahmeverfahren im Vorjahr wurde ja unter anderem aufgrund des Hinweises in den Vorjahresdaten auf ein mögliches Qualitätsdefizit gestartet, und dieser Hinweis wird nun auch im aktuellen Jahr berücksichtigt. Entscheidet man sich alternativ dazu, die Vorjahresdaten im Falle eines Stellungnahmeverfahrens im Vorjahr aus der Betrachtung auszuschließen, so kann man die Wachsamkeit anderweitig erhöhen, indem man im Falle eines Stellungnahmeverfahrens im Vorjahr die Parameter der Entscheidungsregeln (also z. B. α, ζ oder c) verändert.

So eine Anpassung der Parameter der Entscheidungsregeln an das Ergebnis eines Stellungnahmeverfahrens erfordert jedoch weitere technische Festlegungen. Eine einfachere, zielführendere Alternative ist es, den Wunsch, einen Leistungserbringer im Folgejahr nach einem Stellungnahmeverfahren erneut gezielt zu betrachten, während des Stellungnahmeverfahrens separat festzuhalten. Die erneute Betrachtung könnte dann losgekoppelt von quantitativen Überlegungen ausgelöst werden.

Die verschärfte Betrachtung von Leistungserbringern, bei denen im Vorjahr ein Stellungnahmeverfahren stattfand, erweist sich auch im Widerspruch zur optimalen Strategie im Sinne der Entscheidungstheorie (vgl. Abschnitt 5.4.4). Dies bedeutet nicht, dass so eine verschärfte Betrachtung nicht sinnvoll sein kann, sondern liegt darin begründet, dass diese verschärfte Betrachtung (wo es darum geht, wie Leistungserbringer auf die Ergebnisse des Stellungnahmeverfahrens reagieren) ein anderes Ziel hat als die quantitative Auffälligkeitseinstufung im Allgemeinen (wo es darum geht, unbekannte Qualitätsdefizite zu detektieren). Dies ist ein weiteres Argument dafür, erneute Betrachtungen in Folgejahren nach Stellungnahmeverfahren unabhängig von der allgemeinen quantitativen Einstufungsmethodik zu betrachten.

Im Folgenden wird stets davon ausgegangen, dass es für die quantitative Einstufung unerheblich ist, ob im Vorjahr bereits ein Stellungnahmeverfahren stattgefunden hat.

5.4.5.2 Probleme bei zeitlichen Veränderungen in Daten und Indikatoren

Bei Änderungen in der Spezifikation oder in der Definition von Indikatoren muss man prüfen, ob die Berücksichtigung von Vorjahresergebnissen bei der Bestimmung des Hinweises auf ein Qualitätsdefizit Sinn macht. Die meisten der in diesem Abschnitt vorgeschlagenen Methoden nehmen an, dass sich QI-Ergebnisse auch über Mehrjahreszeiträume berechnen lassen.

Außerdem sollte möglichst ein konstanter Referenzbereich gegeben sein. Ändert sich der Referenzbereich, so muss man überlegen, ob man die Vorjahresdaten auch mit dem neuen Referenzbereich auswerten will. Unter Umständen kann dies durchaus legitim sein, sollte aber begründet werden: Die Idee ist ja, die aktuellen QS-Daten im Lichte der Vorjahresdaten zu bewerten. Die Vorjahresdaten können natürlich Evidenz dafür enthalten, dass der Parameter θ außerhalb des neuen Referenzbereichs liegt, auch wenn man dem Leistungserbringer nicht vorwerfen kann, dass der Parameter im letzten Jahr noch nicht im neuen Referenzbereich lag. Schwierig zu interpretieren sind in diesem Kontext allerdings perzentilbasierte Referenzbereiche, die sich jedes Jahr ändern.

5.4.5.3 Perzentilbasierte Referenzwerte

Für sequentielle Verfahren, bei denen Daten von mehr als einem Erfassungsjahr in die Auffälligkeitsklassifikation einfließen, wird vorgeschlagen, den für die Einstufung relevanten Referenzbereich anhand der Daten des aktuellsten Erfassungsjahres – wie oben beschrieben – zu bestimmen und diesen Referenzbereich für die gesamte sequentielle Entscheidungsregel anzuwenden. Dies kann dazu führen, dass unter Umständen Vorjahresergebnisse quantitativ auffällig sind, die zum Referenzbereich des Vorjahres unauffällig waren, und dass in der Regel mehr als die durch das Perzentil festgelegte Anzahl an Leistungserbringerergebnissen auffällig wird. Dies folgt daraus, dass die Vorjahresdaten dazu genutzt werden, stärkere Evidenz dafür zu generieren, ob ein Leistungserbringerergebnis innerhalb des *aktuellen* Referenzbereichs liegt oder nicht.

5.4.5.4 Risikoadjustierung

Die generellen Ideen dieses Abschnitts lassen sich auch auf risikoadjustierte Indikatoren anwenden. Bei der Untersuchung der Auslösewahrscheinlichkeiten ist jedoch das Risikoprofil des betrachteten Leistungserbringers ein weiterer Parameter, der modelliert werden muss.

Viele risikoadjustierte Indikatoren besitzen einen perzentilbasierten Referenzbereich. Außerdem wird oft die Risikoadjustierung regelmäßig überarbeitet, wodurch sich die Rechenregeln ändern. In diesen Fällen sind die Bemerkungen aus dem vorangegangenen Abschnitt 5.4.5.2 zu beachten.

5.4.6 Fazit

In diesem Abschnitt wurden verschiedene Methoden vorgestellt, wie man auf der Basis zweier aufeinanderfolgender Erfassungsjahre zu einer quantitativen Auffälligkeitseinstufung eines Leistungserbringers kommen kann. Die Methoden wurden anhand des Verhaltens der Auslösewahrscheinlichkeit als Funktion von Fallzahl und Kompetenzparameter miteinander verglichen. Unter den vorgestellten Methoden haben die statistisch signifikante Auffälligkeit auf zwei Jahre und die GLR-Binom-Methode die besten Eigenschaften. Die Auslösewahrscheinlichkeit oszilliert schwächer als bei den anderen Methoden als Funktion von der Fallzahl, und Auslösewahrscheinlichkeit steigt stärker mit dem Kompetenzparameter θ an. Ein weiterer Vergleich fand im Rahmen eines konkret operationalisierten Entscheidungsdiagramms anhand des erwarteten Verlusts statt, hier konnten die Methoden mit der optimalen Entscheidungsstrategie verglichen werden: wiederum schnitten die GLR-Binom-Methode und die statistisch signifikante Auffälligkeit auf zwei Jahren am besten ab.

Die statistisch signifikante Auffälligkeit über zwei Jahre hat den zusätzlichen Vorteil, dass die verwendeten Konzepte so oder so ähnlich bereits in der externen stationären Qualitätssicherung etabliert sind. Eine Prüfung der statistischen Signifikanz findet beispielsweise bereits im Verfahren plan. QI statt (IQTIG 2016). Dadurch kann man erwarten, dass bei Landesgeschäftsstellen für Qualitätssicherung und Leistungserbringern ein Grundverständnis über die Rolle des Parameters α vorhanden ist (auch wenn sich die Interpretation im Kontext der erweiterten Datengrundlage ändert, wie im Abschnitt 5.4.2.4 erläutert). Hingegen ist das GLR-Binom-Verfahren relativ technisch, und der Parameter c ist schwer interpretierbar.

5.5 Illustration der Auswertungsmethodik

Die vorgeschlagene statistische Auswertungsmethodik wird im Folgenden beispielhaft anhand eines an den QI 54003 aus dem QS-Verfahren *Hüftendoprothesenversorgung* angelehnten Raten-indikator für einen Leistungserbringer dargestellt.

Zähler, Nenner und Referenzbereich des Qualitätsindikators 54003 sind in Tabelle 13 dargestellt.

Tabelle 13: Rechenregeln eines Beispielindikators, angelehnt an den Qualitätsindikator 54003 „Präoperative Verweildauer bei endoprothetischer Versorgung einer hüftgelenknahen Femurfraktur“

Zähler	Eingriffe, bei denen die Operation später als 48 Stunden nach der Aufnahme erfolgte
Nenner	Alle Eingriffe zur endoprothetischen Versorgung einer hüftgelenknahen Femurfraktur
Referenzbereich	≤ 15 %

Gegenüber der bisherigen Methode zur Festelltung einer sog. rechnerischen Auffälligkeit soll zukünftig auf Basis der Beobachtungen von „Anzahl Fälle im Zähler“ und „Anzahl Fälle im Nenner“ eine *Wahrscheinlichkeit* dafür ermittelt werden, dass die wahre, nicht direkt beobachtbare Rate eines Leistungserbringers innerhalb des Referenzbereichs liegt.

Damit soll dem Sachverhalt Rechnung getragen werden, dass gleiche rechnerische Ergebnisse „Anzahl Fälle in Zähler durch Anzahl Fälle in Nenner“ von Leistungserbringern mit unterschiedlicher Nenner-Fallzahl nicht gleichbedeutend sind; Ergebnisse mit größerer Nenner-Fallzahl liefern nämlich stärkere statistische Evidenz dafür, dass die zugrunde liegende Rate eines Leistungserbringers im Qualitätsindikator inner- bzw. außerhalb des Referenzbereichs liegt. Beispielsweise überschreiten die Ergebnisse 50/100 und 1/2 beide den Referenzbereich ≤ 15 % um 35 %-Punkte. Die Evidenz für ein Qualitätsdefizit bei einer Fallzahl von 100 ist jedoch größer, als bei einer Fallzahl von 2, bei der das Ergebnis auch allein aufgrund zufälliger Variation zustande gekommen sein kann.

Daher bezieht sich die vorgeschlagene Einstufungsmethodik nicht auf der beobachteten Rate, welche auch aufgrund stochastischer Einflüsse variiert, sondern bezieht sich auf die wahre, zugrunde liegende Rate. Anhand des beobachteten Indikatorergebnisses kann, unter Berücksichtigung stochastischer Einflüsse, die Wahrscheinlichkeit dafür berechnet werden, dass diese zugrunde liegende Rate innerhalb des Referenzbereichs liegt.

Nach der vorgeschlagenen Auswertungsmethodik wird für einen Leistungserbringer im ersten Schritt die Wahrscheinlichkeit dafür berechnet, dass die zugrunde liegende Rate im betrachteten Indikator innerhalb des Referenzbereichs liegt. Illustriert wird dies im Folgenden anhand eines hypothetischen Leistungserbringers und dem oben genannten Qualitätsindikator mit:

$$J = 20; o = 5$$

Bei 5 von insgesamt 20 behandelten Fällen im aktuellen Erfassungsjahr erfolgte demnach die Operation später als 48 Stunden nach Aufnahme. Die beobachtete Rate des Leistungserbringers liegt bei o/J = 25 %. Um Rückschlüsse auf die zugrunde liegende, nicht direkt beobachtbare Rate des Leistungserbringers zu ziehen, wird auf Basis der beobachteten Daten und mittels des oben beschriebenen bayesianischen Modells die sogenannte A-Posteriori-Wahrscheinlichkeit der zugrundeliegenden Rate θ berechnet. Die Verteilung der Wahrscheinlichkeit für den oben betrachteten Fall in Abbildung 30 dargestellt.

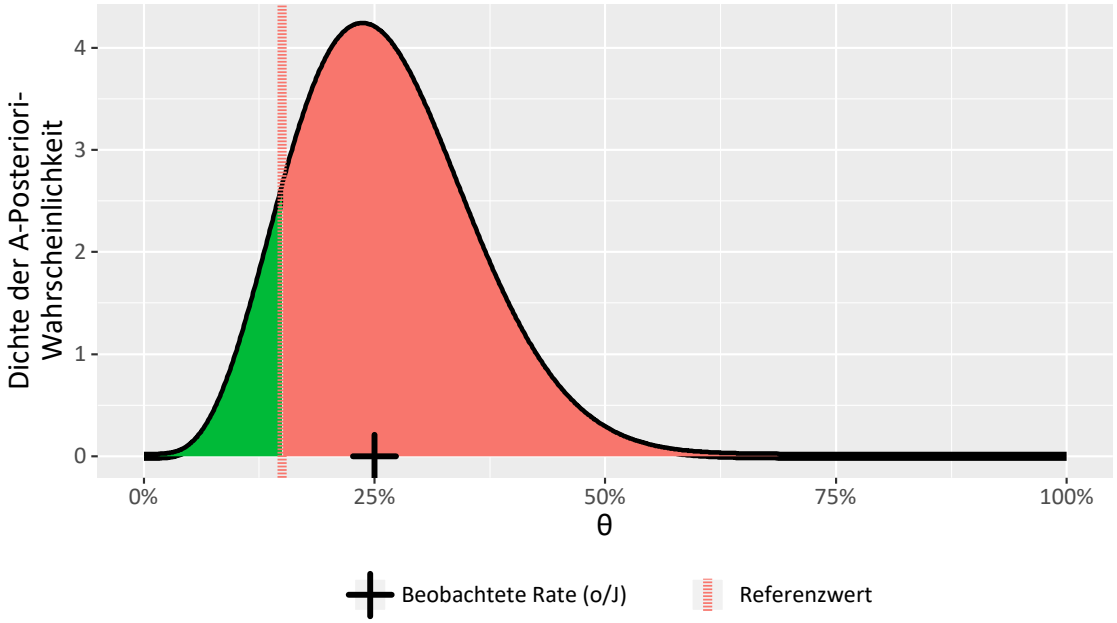

Abbildung 30: Dichte der A-posteriori-Wahrscheinlichkeit für die zugrunde liegende Rate ϑ bei einem Leistungserbringer mit $J = 20$ Fällen im Nenner und $o = 5$ Fällen im Zähler. Anhand des Referenzwerts von 15 % wird diese Dichte in Dichte für Werte innerhalb und außerhalb des Referenzbereiches unterteilt.

Die grüne Fläche kennzeichnet dabei die Dichte der Verteilung innerhalb des Referenzbereichs, die rote Fläche jene außerhalb des Referenzbereichs. Insgesamt summieren die beiden Flächen zu dem Wert 1. Die A-posteriori-Wahrscheinlichkeit dafür, dass die zugrunde liegende Rate θ innerhalb des Referenzbereichs liegt, kann als Fläche der Dichte innerhalb des Referenzbereichs ausgedrückt werden und liegt in diesem Fall bei 11 %. Unterschreitet diese Wahrscheinlichkeit den vorgegebenen Schwellenwert α (Signifikanzniveau), wird das Leistungserbringerergebnis als quantitativ auffällig eingestuft, weil die Wahrscheinlichkeit dafür, dass die zugrunde liegende Rate des Leistungserbringers im Referenzbereich liegt, unwahrscheinlicher ist als α. Ist dies nicht der Fall, wird das Indikatorergebnis als quantitativ unauffällig eingestuft.

Sofern die Einstufungsmethodik die Betrachtung von zwei Erfassungsjahren einschließt, wird in einem zweiten Schritt auf Basis der Daten des aktuellen Erfassungsjahres und des Vorjahres die Wahrscheinlichkeit dafür berechnet, dass die zugrunde liegende Rate *innerhalb* des Referenzbereichs liegt, da der Punktschätzer für das aktuelle Erfassungsjahr außerhalb des Referenzbereichs liegt. Unter der Annahme, dass sich die wahre Rate des Leistungserbringers zwischen den Jahren nicht ändert, erlaubt die Vergrößerung der Datenbasis eine präzisere statistische Einschätzung, ob die wahre Rate innerhalb des Referenzbereichs liegt. Wurde für den Leistungserbringer im Vorjahr dasselbe Ergebnis wie im aktuellen Erfassungsjahr, d. h. $5/20 = 25$ % beobachtet, liegt die auf Basis der gepoolten Daten, d. h. $J = 20 + 20 = 40$ und $o = 5 + 5 = 10$ berechnete Wahrscheinlichkeit dafür, dass die zugrunde liegende Rate im Referenzbereich liegt, bei 0,9 % und ist damit deutlich niedriger, als die nur auf Basis der Daten des aktuellen Erfassungsjahres berechnete Wahrscheinlichkeit. Unterschreitet diese Wahrscheinlichkeit das Signifikanzniveau α, wird das Leistungserbringerergebnis für den QI als quantitativ auffällig eingestuft. Eine interaktive Version der Abbildung, in der sowohl o, J als auch der Referenzwert sowie

das Signifikanzniveau verändert werden können, ist unter https://iqtig.shinyapps.io/illustration_einstufungsmethodik/ zu finden.

5.6 Zusammenfassung und Empfehlungen

In diesem Kapitel wurde ein Rahmenkonzept für die statistische Auswertungsmethodik von Qualitätsindikatoren präsentiert. Ein solches statistisches Rahmenkonzept existiert bisher nicht für die DeQS-RL, ist jedoch eine notwendige Voraussetzung für eine differenzierte Diskussion über die Auswertung von Qualitätsindikatoren. Ein wichtiger Bestandteil des Rahmenkonzepts ist, dass für jeden Indikator bei der Entwicklung entschieden wird, ob der Indikator nach einer analytischen oder enumerativen Herangehensweise ausgewertet werden soll. Der Großteil der Indikatoren in der DeQS-RL ist aus methodischer Sicht des IQTIG nach der analytischen Herangehensweise auszuwerten. Das bedeutet, dass bei der Berechnung und Bewertung der Leistungserbringerergebnisse dieser Indikatoren statistische Unsicherheit zu berücksichtigen ist, indem zwischen dem beobachteten Indikatorwert des Leistungserbringers und dem zugrunde liegenden Parameter des Leistungserbringers (in Abschnitt 5.1.1 auch als Kompetenzparameter bezeichnet) unterschieden wird. Dies ist ein klarer Unterschied zu den Festlegungen in § 10 QSKH-RL, welche die Berücksichtigung statistischer Unsicherheit beim rechnerischen Teil des Bewertungsprozesses (die rechnerische Auffälligkeitseinstufung) explizit ausschließen. Insbesondere auch für die Effizienzsteigerung des Verfahrens ist jedoch die Berücksichtigung statistischer Unsicherheit bei analytisch auszuwertenden Indikatoren notwendig, wie sowohl empirisch in Abschnitt 2.6 (Aufwandsanalyse/Hintergrund) als auch theoretisch in Abschnitt 5.3 gezeigt wird. Eine zentrale Frage, der in diesem Kapitel nachgegangen wird, ist, wie genau statistische Unsicherheit bei der Bewertung der Leistungserbringerergebnisse Berücksichtigung finden soll.

Abschnitt 5.2 stellt aus statistischer Sicht dar, dass der Bewertungsprozess aus zwei Schritten besteht, die für analytisch auszuwertende Indikatoren im Rahmen eines *Entscheidungsproblems unter Unsicherheit* aufgefasst werden können. Der erste Schritt besteht aus einem rein empirischen Vorgehen, bei dem der Begriff „hinreichender Hinweis für ein Qualitätsdefizit“ aus Kapitel 4 als binäre Klassifikation der Leistungserbringerergebnisse operationalisiert wird. Der zweite Schritt besteht in einer fachlichen Bewertung dieser Klassifikation in Form eines Stellungnahmeverfahrens. Für die Klassifikation im ersten Schritt wird in diesem Kapitel der Überbegriff der „quantitativen Auffälligkeitseinstufung“ eingeführt, der konkrete Vorgehensweisen wie die Einstufung in „rechnerische Auffälligkeit“ nach QSKH-RL oder die Einstufung in „statistische Auffälligkeit“ der plan. QI-RL umfasst. Im Rahmen des Entscheidungsproblems müssen zunächst geeignete Kriterien aufgestellt werden, welche die Klassifikation optimieren soll, um festzulegen, welche konkrete Vorgehensweise für die quantitative Auffälligkeitseinstufung zu wählen ist. Die Festlegung dieser Kriterien entspricht der expliziten Festlegung, wie man – aus statistischer Sicht – zu einer Steigerung der Effizienz des gesamten Bewertungsprozesses kommen kann. Im Kapitel werden zwei Vorgehensweisen unterschieden:

- *statistisch signifikante Auffälligkeitseinstufung*: Betrachtet wird die statistisch signifikante Abweichung des Kompetenzparameters des Leistungserbringers vom Referenzbereich. Ziel ist,

anhand der beobachteten Ergebnisse korrekt zu klassifizieren, ob der zugrunde liegende Kompetenzparameter innerhalb oder außerhalb des Referenzbereiches liegt. Dabei spielen die Größe der Abweichung oder die Konsequenzen der Abweichung des Kompetenzparameters keine Rolle – selbst minimale Abweichungen vom Referenzbereich führen zur quantitativen Auffälligkeit.

- *statistisch relevante Auffälligkeitseinstufung*: Das Ausmaß der Abweichung des Kompetenzparameters des Leistungserbringers vom Referenzbereich wird in Bezug zu der Anzahl an betroffenen Patientinnen und Patienten gesetzt. Somit werden die statistischen Hinweise auf ein Qualitätsdefizit in betroffenen Patientinnen und Patienten gemessen.

Jede dieser beiden Vorgehensweisen wird in Abschnitt 5.3.1 theoretisch als Optimierer einer entsprechenden Verlustfunktion in einem bayesianischen Rahmen hergeleitet. Eine Konsequenz der Berücksichtigung statistischer Unsicherheit ist, dass belastbare Aussagen über Ergebnisse von Leistungserbringern mit sehr kleinen Fallzahlen – unabhängig davon, auf welche Art die Unsicherheit berücksichtigt wird (d. h. bei beiden Vorgehensweisen) – schwierig sind, da selten genug statistische Evidenz generiert wird, um von einem hinreichenden Hinweis auf ein Qualitätsdefizit zu sprechen. Während es bei der statistisch signifikanten Einstufung – je nach Wahl des Referenzwerts und des Signifikanzniveaus – für Leistungserbringer meist auch bei sehr kleinen Fallzahlen möglich ist, quantitativ auffällig zu werden,[53] ist eine quantitative Auffälligkeit bei der statistisch relevanten Einstufung erst ab einer gewissen Fallzahl möglich, da diese Vorgehensweise danach geht, wie viele Patientinnen und Patienten von einem möglichen Qualitätsdefizit betroffen sind. Zwar mag dies richtig im vereinfachten Rahmen einer Aufwand-Nutzen-Abwägung des Bewertungsprozesses sein, ist aber kaum vereinbar mit den Zielen der externen Qualitätssicherung nach § 135a Abs. 1 SGB V und könnte unerwünschte, nicht in der Verlustfunktion abgebildete Nebeneffekte haben. Somit müsste man für diese Gruppe zu anderen Mitteln der Bewertung greifen, wie z. B. Peer-Review-Verfahren für eine Stichprobe der nicht berücksichtigten Leistungserbringer. Dies wiederum führt zu einer neuen strukturellen Heterogenität des Bewertungsverfahrens, weil nicht mehr die gleiche Evidenzgrundlage für die Bewertung der Leistungserbringer gesichert ist.

Untersuchungen im Rahmen einer Sensitivitäts- und Spezifitätsanalyse für die Klassifikation des Kompetenzparameters im Abschnitt 5.3.3 zeigen, dass die statistisch signifikante Auffälligkeitseinstufung bei einem zugrunde liegenden Kompetenzparameter, der dem Referenzwert entspricht (die schwierigste Situation für die Klassifikation), einen konstanten Fehler 1. Art besitzt. Dagegen hat die statistisch relevante Auffälligkeitseinstufung in dieser Situation einen mit der Fallzahl ansteigenden Fehler 1. Art. Das bedeutet, dass die Schwellenwerte für eine quantitative Auffälligkeit für größere Leistungserbringer sehr nahe am Referenzwert liegen können und es, bei gleichem zugrunde liegendem Kompetenzparameter, tendenziell schneller zu einer fachlichen Bewertung kommt, als bei kleinen Leistungserbringern. Liegt der Kompetenzparameter etwas weiter vom Referenzwert entfernt, haben beide Methoden geringe Fehler 1. Art. Die

[53] Bei der Wahl von sehr kleinen Signifikanzniveaus kann es aber selbst bei der statistisch signifikanten Auffälligkeitseinstufung passieren, dass ein Leistungserbringer mit einer sehr kleinen Fallzahl nicht auffällig werden kann. Als Beispiel: 1/1 ist nicht auffällig, wenn das Signifikanzniveau $\alpha = 0{,}01$ ist.

Power beider Methoden, d. h. ihre Eigenschaft, Abweichungen des zugrunde liegenden Kompetenzparameters vom Referenzbereich zu detektieren, ist stark fallzahlabhängig und unterscheidet sich je nach Größe der Abweichung. Die bisher in der QSKH-RL verwendete „rechnerische Auffälligkeitseinstufung" kann im analytischen Kontext als eine hoch-sensitive, aber nur sehr niedrig-spezifische Methode angesehen werden, was auch durch die Analysen im Abschnitt 2.6 (Aufwandsanalyse) empirisch belegt wird. Dies führt dann zu einer hohen Menge an Auffälligkeiten, die sich im Stellungnahmeverfahren oft als unbegründet herausstellen.

Jede Entscheidung für eine Methodik zur quantitativen Auffälligkeitseinstufung enthält daher eine Abwägung zwischen Sensitivität und Spezifität der Methode für die Klassifikation des Kompetenzparameters anhand der Konsequenzen. Um diese Abwägung zu systematisieren, ist es notwendig, sowohl Sensitivität als auch Spezifität in einer einheitlichen Metrik zu kombinieren. Hierfür wurde im Abschnitt 5.2 ein Verlustfunktionsansatz im Rahmen des vorliegenden Entscheidungsproblems präsentiert, der in einem vereinfachten Aufwand-Nutzen-Rahmen erlaubt, strukturiert über diese Abwägungen zu diskutieren und Annahmen transparent zu machen. Grundsätzlich fordert der Verlustfunktionsansatz der vorangegangenen Abschnitte eine explizite Festlegung des Aufwandsverhältnisses von Stellungnahmeverfahren gegenüber nicht entdeckten Qualitätsdefiziten. Die Komplexität des Bewertungsprozesses und die vielen Akteure mit unterschiedlichen Aufwandsperspektiven stellen hier eine fast unüberwindbare Herausforderung dar. Trotzdem bietet der entscheidungstheoretische Rahmen mit seinen Vereinfachungen eine Offenlegung der Abwägungen und kann eine Stütze bieten, um z. B. die Schwellenwerte der jeweiligen Einstufungsverfahren zu interpretieren, welche festlegen, ab wann von einem hinreichenden Hinweis auf ein Qualitätsdefizit ausgegangen werden kann. Der vorgestellte entscheidungstheoretische Rahmen erlaubt auch weiterführende Analysen, z. B. zu den Konsequenzen von Fehlklassifikationen[54] bei den interessierenden Ereignissen. Dadurch bietet er prinzipiell auch die Möglichkeit, bei der quantitativen Auffälligkeitseinstufung bereits die fachliche Prüfung im zweiten Schritt des Bewertungsprozesses zu berücksichtigen (Raats und Moors 2003). Diese Überlegungen sind eng mit der Methodik für die Festlegung von Referenzwerten bei Qualitätsindikatoren verknüpft, weil sich u. a.die Frage stellt, ob die jetzigen Referenzwerte bereits Toleranz für besondere Versorgungssituationen enthalten, die nun zusätzlich als stochastische Einflüsse berücksichtigt werden.

Insgesamt zeigen die Analysen, dass die korrekte Klassifikation kleiner Leistungserbringer eine besondere Herausforderung darstellt, die entweder nur über eine Erweiterung der Datenbasis oder mittels nicht statistischer Methoden zu lösen ist. Hier zeigt der in Abschnitt 5.4.3 gebrachte theoriegeleitete Vergleich verschiedener, aus der statistischen Prozesskontrolle inspirierter 2-Jahres-Einstufungsmethoden, dass diese Methoden tatsächlich bessere Klassifikationsergebnisse erreichen, als die 1-Jahres-Einstufungsmethoden.

[54] Mit „Fehlklassifikation" ist hier die folgende Situation gemeint: nach den Rechenregeln des Indikators wird das durch den Indikator betrachtete Ereignis bei einem Fall als unerwünscht kategorisiert, dieses stellt sich jedoch nach der fachlichen Prüfung als besondere Versorgungssituation heraus. Auch die umgekehrte Situation, d. h. ein Ereignis wird fälschlicherweise nicht als unerwünscht kategorisiert, führt zu einer Fehlklassifikation. Prinzipiell können Fehler bei der Datenerhebung des Falls zu Fehlklassifikationen führen.

Konkrete Vorschläge für die statistische Methodik des Bewertungskonzepts: Auslösung von Stellungnahmen

- Die Kriterien für den Bewertungsprozess in Kapitel 4 legen nahe, dass es bei der Bewertung der Leistungserbringerergebnisse bei einem Qualitätsindikator um eine Klassifikation des Kompetenzparameters des Leistungserbringers im Vergleich zum Referenzbereich geht, welche in der Natur des Vergleichs unabhängig von Fallzahl und der Größe der Abweichung vom Referenzwert ist. Aus diesem Grund wird empfohlen, die Einstufung nach der statistisch signifikanten Auffälligkeitseinstufung vorzunehmen. Beim inferenzstatistischen Vorgehen dieser Auffälligkeitseinstufung manifestiert sich dann Fallzahl und Größe der Abweichung vom Referenzwert in den beobachteten Daten als unterschiedliche Grade von empirischer Evidenz für eine Abweichung vom Referenzwert.
- Da bayesianische Inferenz im Gegensatz zu frequentistischer Inferenz einen deutlich flexibleren Rahmen bildet, um Auffälligkeitseinstufungen auch für komplexe Indikatortypen (wie z. B. bei Patientenbefragungen) vorzunehmen, empfiehlt das IQTIG für die DeQS-RL die bayesianische Version der statistisch signifikanten Auffälligkeit (d. h. $S_{\text{stat.sig.bayes}}$, siehe Abschnitt 5.3.1.2) zu verwenden, um einen einheitlichen statistischen Inferenzrahmen für die Auswertungen zu ermöglichen. Dieser Vorschlag entspricht der Empfehlung der Präsidiumskommission der Statistischen Gesellschaften des amerikanischen *Centers for Medicare and Medicaid Services* (CMS) (Ash et al. 2012), hierarchische bayesianische Modelle zu verwenden. Mit der Wahl der Jeffrey-Prior sind die in diesem Kapitel erarbeiteten Vorschläge analytisch kongruent mit der in der plan. QI-RL verwendeten Methode zur Auffälligkeitseinstufung bei Rateindikatoren und risikoadjustierten Indikatoren. Komplexere Überlegungen zur Risikoadjustierung, Einfluss von Fehlklassifikationen, Unsicherheitsfortpflanzung bei der Verwendung von mehreren Datenquellen[55] können nur kohärent im bayesianischen Rahmen vorgenommen werden. Auch bietet der bayesianische Ansatz, wie in Abschnitt 5.3.1 gezeigt, einen Rahmen, um optimale Entscheidungsstrategien für die Auffälligkeitseinstufungsmethode in diesen komplexen Situationen zu analysieren.
- Für belastbare Aussagen bei Leistungserbringern mit kleinen Fallzahlen ist die Datengrundlage eines Erfassungsjahres häufig nicht ausreichend. Es besteht somit die Gefahr, dass mögliche Qualitätsdefizite nicht detektiert werden können. In diesen Fällen empfiehlt das IQTIG, die Datengrundlage zu erweitern und beispielsweise zwei Erfassungsjahre gemeinsam für die Bewertung heranzuziehen. Konkret empfiehlt das IQTIG die bayesianische Version der statistisch signifikanten Auffälligkeitseinstufung basierend auf den Daten zweier Erfassungsjahre ($S_{\text{stat.sig.bayes2}}$, siehe Abschnitt 5.4.2.3), unter der Einschränkung, dass die Methodik zur fachlichen Prüfung und Neuberechnung diese Erweiterung der Datenbasis auf zwei Erfassungsjahre erlaubt oder sie entsprechend erweitert werden kann (vergleiche Empfehlungen in Kapitel 11). Um Erfahrungen mit dem Verfahren zu bekommen und um die in Abschnitt 5.4.5 erwähnten Probleme zu lösen, empfiehlt das IQTIG vor einer flächendeckenden Umsetzung, diese Methodik zunächst in einem direkten QS-Verfahren einzusetzen.

[55] Wie z. B. bei Risikoadjustierungsmodellen, Risikostatistik bzw. Sozialdaten.

- Sowohl $S_{\text{stat.sig.bayes}}$ als auch $S_{\text{stat.sig.bayes2}}$ hängen vom Schwellenwert α ab, welcher als Signifikanzniveau interpretiert werden kann. Dieser Tuning-Parameter stellt zusammen mit dem Referenzwert des Qualitätsindikators die Stellschrauben der Operationalisierung des hinreichenden Hinweises auf ein Qualitätsdefizit im Indikator dar. Diese neue Stellschraube ist eine Konsequenz davon, dass im Rahmen einer analytischen Herangehensweise Unsicherheit bei der Klassifikation der Leistungserbringerergebnisse berücksichtigt wird, und legt fest, ab wann unter Berücksichtigung von Unsicherheit von hinreichender statistischen Evidenz für eine Abweichung vom Referenzbereich ausgegangen werden kann. Somit ist α sorgfältig und behutsam zu wählen: Möchte man eine einheitliche Grundlage für die Operationalisierung des Begriffs „hinreichender Hinweis auf ein Qualitätsdefizit“ über alle QS-Verfahren haben, dann ist *ein* Wert von α für alle QS-Verfahren festzulegen. Möchte man dagegen berücksichtigen, dass unterschiedliche Konsequenzen, Ressourcen und Verfahrenszwecke eine Rolle spielen, kann es ggf. eine unterschiedliche Wahl von α für die quantitative Auffälligkeitseinstufungsmethode geben. Dabei ist zu beachten, dass die 2-Jahres-Einstufung einen leicht höheren Fehler 1. Art hat, als das nominal vorgegebene α, sodass das Erreichen des nominellen Niveaus nur durch die Verwendung eines $\alpha^* < \alpha$ geschieht.
- Das IQTIG empfiehlt für die Wahl des Signifikanzniveaus für die statistische signifikante Auffälligkeitseinstufung einen der folgenden Werte, die auch über eine Abwägung zwischen Aufwänden für ein Stellungnahmeverfahren und den Aufwänden für ein nicht entdecktes Qualitätsdefizit hergeleitet werden können, d. h. es besteht die Möglichkeit, die Relevanz und Auswirkungen der im Indikator betrachteten interessierenden Ereignisse mit zu berücksichtigen:
  - Ein nach Aufwand-Nutzen-Überlegungen festgelegtes α, welches die Methodik zur Wahl des Referenzwerts und quantitative Rückmeldungen zur Treffsicherheit des Qualitätsindikators aus den Stellungnahmeverfahren mit berücksichtigt. Beispielsweise enthalten einige der jetzigen Referenzwerte einen gewissen Puffer für Fehlklassifikationen, die später im Stellungnahmeverfahren geklärt werden sollen. Dieser Puffer kompliziert die Aufwand-Nutzen-Überlegungen. Um daher solche Festlegungen adäquat treffen zu können, sind weitere methodische Arbeiten bzgl. der Wahl der Referenzwerte sowie Korrekturen für Fehlklassifikationen bei den entscheidungstheoretischen Überlegungen notwendig.
  - α = 0,05 – dies entspricht der Auffälligkeitseinstufungsmethode der plan. QI-RL und würde zu einer konsistenten Vorgehensweise mit dieser Richtlinie führen.
  - Ein α-Wert, welcher in etwa die gleiche Anzahl an Stellungnahmen produziert, wie sie zzt. beobachtet wird – vgl. Abschnitt 10.4 für eine exemplarische Betrachtung bei den QS-Verfahren *Hüftendoprothesenversorgung* und *Nierentransplantation*. Für QS-Verfahren, bei denen jede rechnerische Auffälligkeit zu einer Stellungnahme geführt hat, würde dies einem α von 0,465 (entsprechend der rechnerischen Auffälligkeitseinstufung) bedeuten. Die konkrete Wahl würde somit vom betrachteten QS-Verfahren abhängen oder müsste als Mittelung verschiedener Verfahren vorgenommen werden.
- Perspektivisch sollte geprüft werden, inwieweit es sinnvoll ist, bei der Wahl des α-Werts zu berücksichtigen, wie oft es bei einem Qualitätsindikator zu Unterschieden zwischen der quantitativen Auffälligkeitseinstufung und der fachlichen Prüfung kommt. Beispielsweise könnte

es sinnvoll sein, die notwendige statistische Evidenz für einen hinreichenden Hinweis auf ein Qualitätsdefizit umso geringer anzusetzen, je größer die Übereinstimmung beider Verfahren ist. Dafür sollen im Rahmen des vorgeschlagenen Stellungnahmeverfahrens Daten dazu gesammelt werden, wie groß diese Übereinstimmung bzw. Abweichung auf QI-Ebene ist. Anschließend sollte geprüft werden, ob die oben dargelegten Vorschläge für die Wahl von α um diese Komponente auf QI- oder QS-Verfahrensebene erweitert werden sollen (siehe Kapitel 11).

- Obwohl für die quantitative Auffälligkeitseinstufung unterschiedliche Signifikanzniveaus und Datengrundlagen gewählt werden können, sollte das Public Reporting der Ergebnisse eines Erfassungsjahres bei den Unsicherheitsintervallen immer ein konsistentes Signifikanzniveau festlegen. Im Rahmen der Qualitätsberichte der Krankenhäuser wäre dies zzt. $\alpha = 0{,}025$, um damit zweiseitige 95 %-Unsicherheitsintervalle für die Ergebnisse eines Erfassungsjahres zu bekommen.[56]

Eine umsetzungsorientierte Evaluation zu Konsequenzen zur Wahl des Signifikanzniveaus findet sich in Abschnitt 10.3 und daraus abgeleitete Empfehlungen in Abschnitt 11.5.1.

Zusammenfassend ist die Berücksichtigung von Unsicherheit einer der großen Gewinne der vorgestellten statistischen Auswertungsmethodik, weil es ein fachlich fundiertes Konzept darstellt, um das zurzeit verwendete Instrument der Hinweise als Teil des Bewertungsprozesses zu vermeiden. Unsicherheit ist inhärenter Bestandteil des Bewertungsprozesses, jedoch können statistische Methoden helfen, die Konsequenzen von Unsicherheit bei der Bewertung zu standardisieren und zu reduzieren.

[56] Siehe in den Regelungen des Gemeinsamen Bundesausschusses gemäß § 136b Absatz 1 Satz 1 Nummer 3 SGB V über Inhalt, Umfang und Datenformat eines strukturierten Qualitätsberichts für nach § 108 SGB V zugelassene Krankenhäuser (Regelungen zum Qualitätsbericht der Krankenhäuser, Qb-R), Anlage 1: Inhalt, Umfang und Datenformat eines strukturierten Qualitätsberichts für das Berichtsjahr 2018, Abschnitt C-1.2.2 Ergebnisse für Qualitätsindikatoren.

6 Fachliche Bewertung

Durch die Abgabe von Stellungnahmen erhalten Leistungserbringer, deren Indikatorergebnis einen hinreichenden Hinweis auf ein Qualitätsdefizit geliefert hat, die Möglichkeit, besondere Konstellationen geltend zu machen, die das Verfehlen des Referenzbereichs erklären und von ihnen nicht zu verantworten sind, aber in der indikatorbasierten Bewertung nicht berücksichtigt sind (siehe Abschnitt 3.1). Die fachliche Bewertung einer Stellungnahme hat die Funktion, die in der Stellungnahme angeführten Gründe für das Verfehlen des Referenzbereichs zu prüfen und zu beurteilen, ob tatsächlich eine solche besondere, nicht vom Leistungserbringer zu verantwortende Konstellation vorliegt, die den Hinweis aus dem Indikatorergebnis entkräftet. Auf Grundlage der fachlichen Bewertung wird eine abschließende Einstufung des Indikatorergebnisses vorgenommen.

In den folgenden Abschnitten wird zunächst hergeleitet, in welche Kategorien diese Einstufung erfolgen sollte. Das daraus folgende Bewertungsschema ist in Abschnitt 6.1.4 dargestellt. In den Abschnitten 6.2 bis 6.5 werden Methodik, Ablauf und Rahmenbedingungen der fachlichen Bewertung erläutert, mit denen die Einstufung in das Bewertungsschema vorgenommen werden soll. In Abschnitt 6.6 werden praktische Umsetzungsfragen und der erwartete Aufwand für die Umsetzung der Empfehlungen diskutiert.

6.1 Bewertungsschema zur Einstufung der Indikatorergebnisse

Die Ergebnisse der Bewertung im Stellungnahmenverfahren sind der Ausgangspunkt für verschiedene Handlungsanschlüsse. Zum einen werden die Ergebnisse für den Zweck eingesetzt, Rechenschaft über die erreichte Versorgungsqualität abzugeben, im Sinne von *accountability* (siehe Abschnitt 2.1). Zum anderen dienen die Bewertungsergebnisse dem Zweck, die Qualitätsverbesserung der Leistungserbringer anzustoßen, unterstützt durch die Qualitätsförderung externer Stellen.

Das Schema, nach dem die Ergebnisse der fachlichen Bewertung kategorisiert werden, soll diese unterschiedlichen Zwecke und damit einhergehenden unterschiedlichen Adressaten berücksichtigen. Während für den Zweck der *accountability* eine eindeutige, für die Allgemeinheit verständliche Aussage über die erreichte Qualität im Vordergrund steht, ist für den Zweck der Qualitätsförderung eine differenziertere Darstellung anzustreben, die die Auswahl angemessener Fördermaßnahmen erlaubt, indem sie die Gründe für vom Referenzbereich abweichende Indikatorergebnisse benennt, also insbesondere zwischen Datenfehlern und fachlich-inhaltlichen Gründen unterscheidet. Diese Anforderungen an den Detailgrad des Bewertungsschemas sind zunächst gegensätzlich.

Das vom IQTIG im Folgenden vorgeschlagene Bewertungsschema stellt daher einen Kompromiss dar, der die oben genannten Anforderungen berücksichtigt, indem zwei Ebenen verwendet werden. In diesem Schema soll durch die Kategorien auf der ersten Ebene eine klare und allgemeinverständliche Aussage getroffen werden, ob ein Leistungserbringer die an ihn gestellten Anforderungen erfüllt oder nicht. Eine weitere Differenzierung dieser Aussage z. B.

für Public Reporting wird auf dieser Darstellungsebene als nicht sinnvoll beurteilt. So ist es für Patientinnen und Patientinnen wenig hilfreich, zu unterscheiden, ob ein Qualitätsdefizit auf fachliche Probleme, Organisationsprobleme oder Dokumentationsprobleme zurückzuführen ist und mit welcher Methode (Vergleich mit dem Referenzbereich versus fachliche Bewertung) die Aussage getroffen wurde. Daher wird empfohlen, diese Details auf einer zweiten Ebene des Schemas zu dokumentieren.

Außerdem beschränkt sich das Bewertungsschema im Einklang mit der empfohlenen Trennung zwischen Qualitätsbewertung und Qualitätsförderung auf das Ergebnis der abschließenden Qualitätsbewertung[57]. Eine Dokumentation der eingeleiteten Fördermaßnahmen wird zwar ebenfalls empfohlen, soll jedoch separat von der Qualitätsbewertung erfolgen (vgl. Kapitel 7).

6.1.1 Kategorien des Bewertungsschemas

Es wird vorgeschlagen, die bisherigen Bezeichnungen „qualitativ auffällig“ und „qualitativ unauffällig“ durch eindeutigere Begrifflichkeiten zu ersetzen. Gerade das Adjektiv „qualitativ“ birgt ein Risiko für Missverständnisse, da es sich sowohl auf den Prozess der Bewertung als auch auf das Ergebnis beziehen kann: Das Ergebnis des qualitativen (im Gegensatz zum quantitativen) Bewertungsprozesses könnte „unauffällig“ lauten oder die Qualität der Versorgung könnte als Ergebnis des gesamten Bewertungsprozesses (quantitativ und qualitativ) als „unauffällig“ bewertet worden sein. Auch bleibt bei dieser Bezeichnung offen, ob „auffällige“ Qualität positiv oder negativ ist. Eine eindeutige Interpretation der Bewertungsergebnisse ist jedoch eine wichtige Voraussetzung für die angemessene Verwendung der Qualitätsergebnisse z. B. für die öffentliche Berichterstattung oder für Handlungsentscheidungen von Institutionen des Gesundheitswesens (siehe Abschnitt 2.2 und 3.7).
Für die Verwendung im Rahmen von *accountability* soll daher am Ende der fachlichen Bewertung eine der folgenden Bewertungskategorien für das Indikatorergebnis eines Leistungserbringers vergeben werden:

- Kein Hinweis auf Qualitätsdefizit
- Qualitätsdefizit
- Sonstiges

Die Bezeichnung „Qualitätsdefizit“ kennzeichnet dabei allgemeinverständlich das Nicht-Erreichen der Anforderungen, die durch den Referenzbereich des Indikators für erwartbare Qualität repräsentiert werden. Die Bezeichnung „kein Hinweis auf ein Qualitätsdefizit“ soll dementsprechend verwendet werden, wenn nach Abschluss des Stellungnahmeverfahrens kein Anhalt besteht, dass das vom Referenzbereich repräsentierte Soll nicht erreicht wurde.

Aus dem zweischrittigen Vorgehen aus quantitativer Bewertung mittels Qualitätsindikator und Prüfung mittels fachlicher Bewertung können sich für einen Leistungserbringer am Ende des Moduls „Qualitätsbewertung“ die in Abbildung 31 dargestellten Kombinationen von Indikatorergebnis und Ergebnis des Stellungnahmeverfahrens ergeben. Bestätigt sich der

[57] Eine Bewertung der Daten- und Dokumentationsqualität, die im Rahmen von Verfahren zur Datenvalidierung untersucht wurde, sollte anhand eines eigenen Bewertungsschemas erfolgen.

indikatorbasierte hinreichende Hinweis auf ein Qualitätsdefizit im Stellungnahmeverfahren, folgt aus dem Verständnis von Qualitätsindikatoren (vgl. Kapitel 3), dass das mittels des Referenzbereichs für erwartbare Qualität geforderte Soll nicht erreicht wurde. Diese Bestätigung kann entweder durch den Leistungserbringer erfolgen, indem dieser den Verzicht auf eine Stellungnahme erklärt und damit die Gültigkeit der Indikatoraussage anerkennt, oder dadurch, dass in der fachlichen Bewertung einer Stellungnahme keine ausreichenden Gründe anerkannt werden, weshalb der Referenzbereich des Indikators ohne Verantwortung des Leistungserbringers verfehlt wurde. Das bestätigte Unterschreiten des Solls in einem Indikator wird mit der Bezeichnung „Qualitätsdefizit“ gekennzeichnet.

Liegt auf Grundlage des Indikatorergebnisses kein hinreichender Hinweis auf ein Qualitätsdefizit vor oder kann dieser Hinweis im Rahmen des Stellungnahmeverfahrens entkräftet werden, soll die Bewertung „Kein Hinweis auf ein Qualitätsdefizit“ erfolgen.

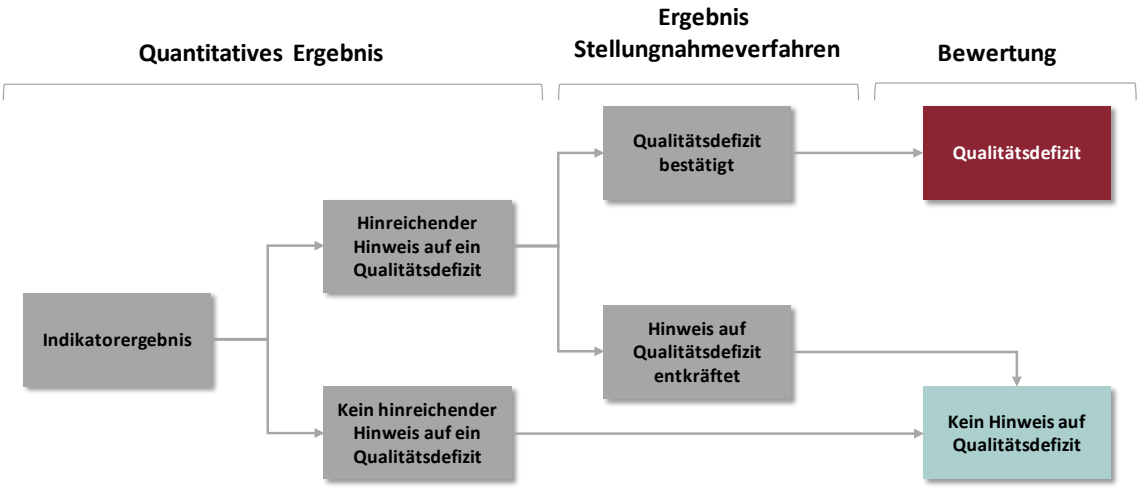

Abbildung 31: Mögliche Kombinationen von quantitativem Indikatorergebnis und Ergebnis des Stellungnahmeverfahrens

Die Kategorie „Sonstiges“ sollte nur in Ausnahmefällen verwendet werden und ist daher immer von der LAG bzw. der Bundesstelle zu begründen, beispielsweise wenn aus organisatorischen Gründen, die nicht vom Leistungserbringer zu verantworten sind, kein Stellungnahmeverfahren durchgeführt werden konnte.

Differenziertere Informationen über die Gründe für die Einstufung in die jeweilige Bewertungskategorie sollen auf der zweiten Darstellungsebene abgebildet werden. Im Folgenden ist erläutert, welche Unterkategorien dazu unterschieden werden sollten.

6.1.2 Differenzierung der Kategorie „kein Hinweis auf Qualitätsdefizit“

Zeigt sich in der auf Indikatorergebnis und ggf. auf dem Stellungnahmeverfahren basierenden Bewertung kein Qualitätsdefizit, so ist zunächst keine weitere Differenzierung erforderlich, da sich keine Maßnahmen der Qualitätsförderung anschließen. Eine weitere Differenzierung dieser Kategorie kann allerdings die Transparenz für die Leistungserbringer erhöhen. Diese können anhand einer genaueren Aufschlüsselung des Bewertungsergebnis auf einen Blick erkennen, wie

ihre Stellungnahme bewertet wurde. Ein weiterer Vorteil besteht darin, dass in der leistungserbringer-übergreifenden Betrachtung deutlich wird, ob es beim betreffenden Qualitätsindikator beispielsweise bei vielen Leistungserbringern zur Entkräftung der Indikatoraussage gekommen ist und ob dies Datenfehlern oder einer besonderen Versorgungskonstellation geschuldet ist. Diese Informationen können zur zielgerichteten Weiterentwicklung der Indikatoren genutzt werden, indem z. B. häufig auftretende Versorgungskonstellationen in der Berechnungsvorschrift des Indikators berücksichtigt werden. Folgende Unterkategorien zur Differenzierung der Kategorie „kein Hinweis auf ein Qualitätsdefizit“ werden daher vorgeschlagen:

- „kein hinreichender Hinweis auf ein Qualitätsdefizit“: Mit dieser Unterkategorie soll dokumentiert werden, dass die quantitative Bewertung anhand des Qualitätsindikators keinen schlüssigen Hinweis auf ein Qualitätsdefizit geliefert hat, da das Indikatorergebnis des Leistungserbringers nicht mit der geforderten statistischen Sicherheit außerhalb des Referenzbereichs lag[58]. Gleichzeitig wird durch Einstufung in diese Unterkategorie dokumentiert, dass kein Stellungnahmeverfahren durchgeführt wurde.

Wurde ein Stellungnahmeverfahren durchgeführt und in diesem die Indikatoraussage entkräftet, sollen die Gründe für die Entkräftung im differenzierten Einstufungsergebnis ablesbar sein (U1 bis U3; siehe Abschnitt 6.4.5):

- „Hinweis auf Qualitätsdefizit entkräftet wegen vom Leistungserbringer nicht zu verantwortender Datenfehler“: Diese Unterkategorie soll anzeigen, dass die Geltendmachung von nicht vom Leistungserbringer zu verantwortenden Datenfehlern[59] für die Entkräftung maßgeblich war.
- „Hinweis auf Qualitätsdefizit entkräftet wegen besonderer Versorgungskonstellation“: Diese Unterkategorie soll dokumentieren, dass die Entkräftung wegen nicht vom Leistungserbringer zu veranwortender, fachlich-inhaltlicher Einflussfaktoren erfolgte.
- „Hinweis auf Qualitätsdefizit entkräftet wegen vom Leistungserbringer nicht zu verantwortender Datenfehler und besonderer Versorgungskonstellation“: Eine Einstufung in diese Unterkategorie soll erfolgen, wenn sowohl Datenfehler als auch fachlich-inhaltliche Einflussfaktoren zur Entkräftung geführt haben. Diese Einstufung soll dabei nicht für jede Konstellation vergeben werden, in der beide Arten von Einflussfaktoren vorgelegen haben, sondern nur für Konstellationen, in denen beide Arten von Einflussfaktoren für die Einstufung (in die Hauptkategorie) *entscheidend* waren. Dies ist dann gegeben, wenn entweder erst die Kombination aus beiden Arten von Gründen zur Entkräftung des Indikatorergebnisses geführt

[58] Aus Gründen der besseren Verständlichkeit werden in diesem Kapitel die Bezeichnungen „hinreichender Hinweis auf ein Qualitätsdefizit“ und „Verfehlen des Referenzbereichs“ synonym verwendet. In beiden Fällen ist gemeint, dass das Indikatorergebnis unter Berücksichtigung von stochastischen Einflüssen den Referenzbereich nicht erreicht, d. h. es ist kein einfacher Größer-kleiner-Vergleich des Indikatorwerts mit dem Referenzbereich gemeint. Details zur Berechnung des indikatorbasierten hinreichenden Hinweises auf ein Qualitätsdefizit sind in Kapitel 5 beschrieben.

[59] Im Bericht wird der Begriff „Datenfehler“ statt „Dokumentationsfehler“ verwendet, um deutlich zu machen, dass in seltenen Fällen auch Fehler in den Daten vorliegen können, die nicht vom Leistungserbringer zu verantworten sind, also keine Fehler des Dokumentationsprozesses darstellen (siehe Abschnitt 6.4.1).

hat oder wenn sowohl Datenfehler alleine als auch fachlich-inhaltliche Faktoren alleine für die Entkräftung hinreichend gewesen wären.

6.1.3 Differenzierung der Kategorie „Qualitätsdefizit“

Wird ein Indikatorergebnis nach Abschluss des Stellungnahmeverfahrens mit „Qualitätsdefizit“ bewertet, sollen die Gründe für die Bewertung transparent sein. Sowohl für den Leistungserbringer als auch für die Stelle, die das Stellungnahmeverfahren durchführt, kann durch die transparente Darstellung der Gründe für die Bewertung deutlich gemacht werden, wie eine Stellungnahme gewertet wurde und mit welchem Fokus Maßnahmen des internen Qualitätsmanagement bzw. externe Maßnahmen der Qualitätsförderung eingesetzt werden sollten. Zudem können in einer leistungserbringerübergreifenden Betrachtung Hinweise zur Datenqualität und Validität des Qualitätsindikators gewonnen werden (siehe oben). Die Dokumentation, welche der Stellungnahmen wegen Nichteinhaltung der formalen Vorgaben nicht fachlich-inhaltlich bewertet werden konnten, liefert darüber hinaus wichtige Hinweise zur Durchführungsevaluation des Stellungnahmeverfahrens (s. a. Kapitel 10).

Differenzierung zwischen Datenfehlern und Defizit in der Versorgungsqualität

In seiner Stellungnahme zum Indikatorergebnis kann ein Leistungserbringer sowohl Datenfehler als auch fachlich-inhaltliche Einflussfaktoren geltend machen. Beide Komponenten können zum Verfehlen des Referenzbereichs des Indikatores beigetragen haben. Für die Ableitung geeigneter Verbesserungsmaßnahmen ist es wichtig, die Gründe für das Verfehlen korrekt zu identifizieren. Außerdem sollen bei der fachlichen Bewertung Defizite der Versorgungsqualität nicht durch angegebene oder tatsächliche Datenfehler verdeckt werden, damit Anreize vermieden werden, die Qualitätsbewertung zu umgehen. Für die Differenzierung zwischen Datenfehlern und Defiziten der Versorgungsqualität werden daher folgende Unterkategorien empfohlen:

- Wenn Dokumentationsfehler, d. h. vom Leistungserbringer zu verantwortende Datenfehler, entscheidend für das Verfehlen des Referenzbereichs des Qualitätsindikators waren, es also in der fachlichen Bewertung ohne diese Dokumentationsfehler nicht zu einer Bewertung als „Qualitätsdefizit“ gekommen wäre, soll die Einstufung „Vom Leistungserbringer zu verantwortende Datenfehler haben eine Fehlmessung verursacht“ (Unterkategorie A1) vergeben werden.

Ist das Verfehlen des Referenzbereichs durch den Leistungserbringer nicht allein durch Dokumentationsfehler zu erklären, so ist zumindest ein Teil dieser Abweichung durch ein Defizit der Versorgungsqualität bedingt. In diesem Fall soll das Bewertungsergebnis widerspiegeln, in welchem Ausmaß die beiden Komponenten Dokumentationsfehler bzw. Versorgungsdefizit jeweils *entscheidend* für die Bewertung als „Qualitätsdefizit“ waren:

- Werden in der fachlichen Bewertung sowohl Dokumentationsfehler als auch Defizite der Versorgungsqualität als bedeutsame Gründe identifiziert, soll dies durch die Unterkategorie „Qualitätsdefizit, teilweise durch vom Leistungserbringer zu verantwortende Datenfehler bedingt“ gekennzeichnet werden (Unterkategorie A2). Diese Unterkategorie soll also nicht jede

Konstellation kennzeichnen, bei der sowohl Datenfehler als auch fachlich-inhaltliche Faktoren zum Verfehlen des Referenzbereiches geführt haben[60], sondern vielmehr solche Konstellationen, bei denen beide Komponenten in relevantem Ausmaß zur Abweichung beigetragen haben.

- Die Unterkategorie „Hinweis auf Qualitätsdefizit im Stellungnahmeverfahren bestätigt oder Verzicht auf Einreichung einer Stellungnahme“ (A3) soll dagegen darstellen, dass für die Bewertung als „Qualitätsdefizit“ fachlich-inhaltliche Gründe führend waren. Dies ist insbesondere dann gegeben, wenn der Leistungserbringer keine Datenfehler geltend gemacht hat oder durch Verzicht auf eine Stellungnahme die Validität der Indikatoraussage anerkannt hat.

Differenzierung zwischen inhaltlichen und formalen Defiziten

Ist wegen der Nichteinhaltung formaler Vorgaben durch einen Leistungserbringer keine fachliche-inhaltliche Bewertung einer Stellungnahme möglich, kann die oben erläuterte Differenzierung der Gründe für ein Qualitätsdefizit nicht vorgenommen werden. Für diesen Fall wird im Bewertungsschema die zusätzliche Unterkategorie „Stellungnahme nicht fristgerecht eingereicht oder Stellungnahme entsprach nicht den formalen Anforderungen“ benötigt. Eine Einstufung in diese Unterkategorie sollte aus den in Abschnitt 6.5.1 erläuterten Gründen als Qualitätsdefizit gewertet werden (Unterkategorie A0). Sie bringt zum Ausdruck, dass der indikatorbasierte Hinweis auf ein Qualitätsdefizit nicht entkräftet wurde, da vom Leistungserbringer keine adäquaten Informationen für eine Entkräftung zur Verfügung gestellt wurden.

6.1.4 Bewertungsschema im Überblick

Tabelle 14 zeigt das vollständige Bewertungsschema im Überblick. Wie bisher sollen auch zukünftig das Bewertungsergebnis auf der oberen Ebene durch Buchstaben gekennzeichnet und die Binnendifferenzierung der Bewertungskategorien durch Ziffern kodiert werden. In Anlehnung an den bisherigen Schlüssel werden für ein Qualitätsdefizit der Buchstabe A, für fehlende Hinweise auf ein Qualitätsdefizit der Buchstabe U und für die Kategorie „Sonstiges“ der Buchstabe S vorgeschlagen. In den Abschnitten 6.2 bis 6.4 dieses Kapitels wird beschrieben, mit welchen Methoden im Rahmen der fachlichen Bewertung eine angemessene Einstufung in diese Kategorien und Ziffern erfolgen kann.

[60] Dies würde andernfalls dazu führen, dass z. B. bereits bei Feststellung eines einzelnen Behandlungsfalls mit Datenfehler die Mischkategorie vergeben würden, obwohl der überwiegende Teil der Abweichung vom Qualitätsziel inhaltlich bedingt war.

Tabelle 14: Bewertungsschema

	Kategorie	Einstufung	Ziffer	Erläuterung
Ergebnis der Qualitätsbewertung	**U**	**Kein Hinweis auf Qualitätsdefizit**	0	Kein hinreichender Hinweis auf ein Qualitätsdefizit
			1	Hinweis auf Qualitätsdefizit entkräftet wegen vom Leistungserbringer nicht zu verantwortender Datenfehler
			2	Hinweis auf Qualitätsdefizit entkräftet wegen vom Leistungserbringer nicht zu verantwortender Datenfehler und besonderer Versorgungskonstellation
			3	Hinweis auf Qualitätsdefizit entkräftet wegen besonderer Versorgungskonstellation
	A	**Qualitätsdefizit**	0	Stellungnahme nicht fristgerecht eingereicht oder Stellungnahme entsprach nicht den formalen Anforderungen
			1	Vom Leistungserbringer zu verantwortende Datenfehler haben eine Fehlmessung verursacht
			2	Qualitätsdefizit, teilweise durch vom Leistungserbringer zu verantwortende Datenfehler bedingt
			3	Hinweis auf Qualitätsdefizit im Stellungnahmeverfahren bestätigt *oder* Verzicht auf Einreichung einer Stellungnahme
keine Qualitätsbewertung	S	**Sonstiges – ohne Bewertung**	-	*[genauer Sachverhalt durch LAG/Bundesstelle anzugeben]*[61]

6.1.5 Empfehlung für die zukünftige Einstufung der Indikatorergebnisse

Im Bewertungsschema gemäß Tabelle 14 wird zwischen „Qualitätsdefizit“ auf der einen Seite und „kein Hinweis auf Qualitätsdefizit“ auf der anderen Seite unterschieden. In diesen Bezeichnungen spiegelt sich wider, dass die Abwesenheit eines hinreichenden Hinweises auf ein Qualitätsdefizit nicht gleichzusetzen ist mit einem Hinweis auf zureichende Qualität. Aufgrund des Einflusses unbekannter oder nicht gemessener Faktoren auf den beobachteten

[61] Der genaue Sachverhalt wird in Textform verpflichtend im Qualitätsbericht des Krankenhauses veröffentlicht.

Indikatorwert ist dieser mit statistischer Unsicherheit verbunden[62] und kann von dem „wahren", zugrunde liegenden Indikatorwert des Leistungserbringers abweichen (IQTIG 2019a: 163 f.). Unter Berücksichtigung dieser Unsicherheit ergibt sich somit noch eine dritte Situation. So kann beispielsweise bei sehr kleinen Fallzahlen und gleichzeitig Indikatorergebnissen nahe der Referenzbereichsgrenze die statistische Unsicherheit so groß sein, dass keine Aussage darüber möglich ist, ob der zugrunde liegende Indikatorwert innerhalb oder außerhalb des Referenzbereichs liegt. In dieser Situation kann weder die Aussage getroffen werden, dass ein hinreichender Hinweis auf zureichende Qualität vorliegt, noch dass ein hinreichender Hinweis auf ein Qualitätsdefizit existiert.

Während die Bezeichnung „Qualitätsdefizit" einen Qualitätsmangel deutlich benennt, bleibt bei der Bezeichnung „kein Hinweis auf Qualitätsdefizit" also offen, ob zureichende Qualität vorliegt oder nur Unsicherheit über die Qualitätsbewertung besteht. Dies ist nicht nur aus Patientensicht, sondern auch aus Sicht der Leistungserbringer verbesserungsbedürftig. Aber auch vor dem Hintergrund der unterschiedlichen Verwendungszwecke der Qualitätsergebnisse, wie dem G-BA-Qualitätsportal oder für die Versorgungssteuerung, ist aus Sicht des IQTIG eine methodisch fundierte Feststellung zureichender Versorgungsqualität unabdingbar. Im Hinblick auf qualitätsverbessernde Maßnahmen nach dem Leitprinzip „Lernen von den Besseren" (siehe Kapitel 7) ist eine Identifizierung von Leistungserbringern mit zureichender Qualität und ggf. zukünftig mit besonders guter Qualität ebenfalls wichtig. Daher wird empfohlen, die Methodik dahingehend weiterzuentwickeln, dass hinreichende Sicherheit über das Erreichen des Solls in einem Indikator (z. B. des Referenzbereichs für erwartbare Qualität) mit einer entsprechenden Bewertungskategorie (z. B. „zureichende Qualität") bezeichnet werden kann.

Die in Kapitel 2 empfohlene biometrische Operationalisierung des hinreichenden Hinweises auf ein Qualitätsdefizit berücksichtigt zunächst nur die Differenzierung zwischen einem hinreichenden Hinweis auf ein Qualitätsdefizit und der Abwesenheit eines solchen Hinweises. Daher ist eine Unterscheidung zwischen Leistungserbringern, deren Indikatorergebnis einen Hinweis auf zureichende Qualität aufweist und Leistungserbringern, deren Indikatorergebnis weder einen Hinweis auf zureichende Qualität noch auf ein Qualitätsdefizit aufweist, mit dieser Methodik zunächst nicht möglich. Um Leistungserbringern, die zureichende Qualität erbringen, dies auch attestieren zu können, wird im Sinne des Rahmenkonzepts für die weitere Verwendung der Ergebnisse z. B. im G-BA-Qualitätsportal oder für Qualitätszu- und -abschläge auch die Weiterentwicklung der biometrischen Methodik um diese Komponente empfohlen. Ansätze zu einer solchen Erweiterung der Methodik werden in Kapitel 2 skizziert.

6.2 Methodik der fachlichen Bewertung

Die fachliche Bewertung der Stellungnahmen soll gemäß den in Kapitel 3 dargestellen methodischen Grundsätzen nach einem möglichst transparenten und objektiven Verfahren erfolgen. Sie hat die Funktion, die in der Stellungnahme angeführten Gründe für das Verfehlen

[62] Wie in Kapitel 5 dargelegt, spielt statistische Unsicherheit bei manchen Indikatoren, wie beispielsweise Strukturindikatoren keine Rolle. Daher kann für Strukturindikatoren die Einhaltung des Referenzbereichs durch einfache Differenzbildung mit der Referenzbereichsgrenze ermittelt und in diesem Fall theoretisch schon jetzt ohne Weiteres zureichende Qualität attestiert werden.

des Referenzbereichs zu prüfen und zu beurteilen, ob eine besondere Konstellation vorliegt, die den Hinweis aus dem Indikatorergebnis entkräftet. Unter einer *besonderen Konstellation* wird in diesem Bericht das Vorliegen eines oder mehrerer Einflussfaktoren auf das Indikatorergebnis bei einem Leistungserbringer verstanden, die nicht von diesem zu verantworten sind und die in ihrer Gesamtheit das Verfehlen des Referenzbereichs bei diesem Qualitätsindikator erklären. Die in einer Stellungnahme genannten Gründe müssen das Vorliegen solcher Einflussfaktoren nachvollziehbar darstellen.

Bei der Beurteilung der Stellungnahmen ist zu berücksichtigen, dass manche Einflussfaktoren nur einen Teil der Behandlungsfälle des Leistungserbringers betreffen oder nur für einen Teil der Behandlungsfälle anerkannt werden können. Darüber hinaus kann die Wirkung eines Einflussfaktors auch bezogen auf einen einzelnen Behandlungsfall unterschiedlich groß sein (siehe Abschnitt 6.4.1). Für die fachliche Bewertung der Stellungnahmen reicht es daher nicht aus, zu bewerten, ob überhaupt ein nicht vom Leistungserbringer zu verantwortender Einflussfaktor bei der Messung vorlag. Es muss vielmehr beurteilt werden, ob das Ausmaß aller in der Stellungnahme zu einem Indikator angeführten und als nicht vom Leistungserbringer zu verantwortend anerkannten Einflussfaktoren in ihrer Gesamtheit groß genug war, um eine besondere Konstellation zu begründen. Dies erfordert eine Zusammenführung und Gewichtung der verschiedenen Einflussfaktoren, damit am Ende des Bewertungsprozesses eine eindeutige Qualitätsaussage je Indikatorergebnis eines Leistungserbringers getroffen werden kann. Gemäß den Ausführungen in Abschnitt 3.4 kann diese Gesamtbeurteilung der Einflussfaktoren und der zugehörigen Begründungen in der Stellungnahme grundsätzlich auf impliziten oder expliziten Regeln basieren.

Eine Beurteilung anhand impliziter Regeln würde bedeuten, dass durch die LAG bzw. die Bundesstelle und die beratenden Expertinnen und Experten durch individuelle Einschätzung abgewogen würde, wie nachvollziehbar die in der Stellungnahme vorgebrachten Gründe sind und wie relevant die dadurch anerkannten Einflussfaktoren für die Bewertung sind. Das Gesamturteil ergäbe sich bei diesem Vorgehen aus der Summe der individuellen Einschätzungen der Bewertenden auf Grundlage ihrer fachlichen Erfahrung und könnte in einem Abstimmungsverfahren ermittelt werden. Dies entspricht größtenteils dem bisherigen Vorgehen. Im Folgenden werden Beurteilungsschritte, die von den Beurteilenden auf Basis solcher impliziter Regeln vorgenommen werden, vereinfachend als „heuristische Beurteilung“ bezeichnet. Ein Vorgehen anhand expliziter Beurteilungsregeln erfordert dagegen eine möglichst eindeutige Beschreibung der Entscheidungssituation einschließlich der Informationsgrundlage sowie eindeutiger Entscheidungsregeln, welche Gründe zu bewerten sind und auf welche Art und Weise die Bewertung erfolgen muss. Um bei einem solchen Vorgehen eindeutige Regeln aufzustellen, nach welchen Kriterien in den Stellungnahmen vorgebrachte Gründe berücksichtigt werden sollen, bedarf es mindestens zum Teil einer expliziten und quantitativen Gewichtung der als Grund anerkannten Einflussfaktoren. Dazu würden für die Einflussfaktoren vorab ein Algorithmus und quantitative Gewichte festgelegt werden, anhand derer eine Neubewertung unter Berücksichtigung dieser Faktoren durchgeführt würde (z. B. in Form einer Nachberechnung des Indikatorwerts für diesen Leistungserbringer und eines erneuten Vergleichs mit einem ggf. angepassten Referenzbereich).

Wie in Abschnitt 3.4 ausgeführt, sind aus methodischer Sicht bei der fachlichen Bewertung explizite Beurteilungsregeln auf Basis standardisierter Informationen zu bevorzugen, da die Erweiterung der Verwendung von Qualitätsmessungen auf Zwecke der *accountability* eine möglichst hohe Objektivität und Reliabilität der Bewertung erfordert. Eine vollständig explizite und quantitative Bewertung der in den Stellungnahmen geltend gemachten Einflussfaktoren ist jedoch aus mehreren Gründen nicht möglich:

- Für die quantitative Neubewertung eines Indikatorergebnisses müsste für alle im Stellungnahmeverfahren vorgebrachten Einflussfaktoren bestimmbar sein, in welchem Ausmaß sie das Indikatorergebnis verändern. Wie in Abschnitt 6.4.1 erläutert, lässt sich jedoch für manche Faktoren nicht zuverlässig angeben, wie sie bei einer Nachberechnung berücksichtigt werden sollten.
- Um beim Stellungnahmeverfahren standardisiert und einheitlich bei den unterschiedlichen LAG angewendet zu werden, müsste die Gewichtung der möglichen Einflussfaktoren bereits vor Beginn des Stellungnahmeverfahrens festgelegt werden. Es handelt sich aber in der Regel um Einflussfaktoren, die in der Rechenvorschrift des Indikators nicht berücksichtigt sind, weil sie zu diesem Zeitpunkt noch nicht bekannt sind oder weil sie zu selten auftreten. In beiden Fällen kann ein Gewicht nicht oder nicht zuverlässig angegeben werden.
- In den Stellungnahmen werden bisher in der Regel nur Angaben zu den Fällen mit einem interessierenden Ereignis (z. B. Komplikationen oder die Nicht-Durchführung eines geforderten Prozesses) gemacht. Um einen Einflussfaktor bei einer Nachberechnung korrekt zu berücksichtigen, müsste die Ausprägung dieses Einflussfaktors für alle Fälle des Leistungserbringers bekannt sein.
- Die Ergebnisse risikoadjustierter Indikatoren basieren nicht nur auf Informationen eines Leistungserbringers, sondern auch auf dem Risikoprofil der Gesamtpopulation. Falls ein Einflussfaktor im Stellungnahmeverfahren als ein zu berücksichtigender Grund anerkannt wird, so müsste daraufhin für eine exakte Nachberechnung das Risikoadjustierungsmodell angepasst werden. Die Berücksichtigung des zusätzlichen Einflussfaktors kann außerdem dazu führen, dass sich auch die Effekte der bereits vorhandenen Einflussgrößen im neuen Risikoadjustierungsmodell ändern. Dies passiert insbesondere, wenn der zusätzliche Einflussfaktor mit den bereits berücksichtigten Risikofaktoren assoziiert ist (beispielsweise weil eine palliative Behandlungsintention bei Patientinnen und Patienten mit dem Vorliegen schwerer Begleiterkrankungen assoziiert ist). Um das Risikoadjustierungsmodell anzupassen, würden jedoch nicht nur für die Leistungserbringer, die im Stellungnahmeverfahren kontaktiert wurden, sondern für alle Leistungserbringer Daten über den neu zu berücksichtigenden Einflussfaktor benötigt.

Eine Einschätzung des Indikatorwertes einer Einrichtung, der vorgelegen hätte, wenn alle nicht vom Leistungserbringer beeinflussbaren Faktoren berücksichtigt worden wären, ist also nur teilweise möglich[63]. Diese Einschätzung ist entsprechend der in Abschnitt 3.1 hergeleiteten

[63] Eine exakte Nachberechnung wäre eine wichtige Voraussetzung für die Verwendung bereinigter Indikatorergebnisse in vergleichenden Darstellungen wie beispielsweise qualitätsbezogenen Rangreihenfolgen. Für diesen Zweck ist das Stellungnahmeverfahren allerdings ohnehin nicht geeignet, da es nicht alle Leistungserbringer umfasst. Vergleiche von nachberechneten Indikatorergebnissen der

Funktion von Stellungnahmen allerdings auch nur soweit notwendig, dass entschieden werden kann, ob der Indikatorwert die Erfüllung der Qualitätsanforderungen durch einen Leistungserbringer abbildet oder nicht. Das IQTIG empfiehlt daher ein kombiniertes Vorgehen, bei dem Einflussfaktoren, deren Beitrag zum Indikatorergebnis eindeutig angegeben werden kann, in Form einer partiellen Nachberechnung des Indikatorwertes und die übrigen angeführten Einflussfaktoren durch heuristische Beurteilung mit Unterstützung der Expertinnen und Experten berücksichtigt werden.

6.3 Ablauf der fachlichen Bewertung

Abbildung 32 gibt einen Überblick über den empfohlenen Ablauf der fachlichen Bewertung.

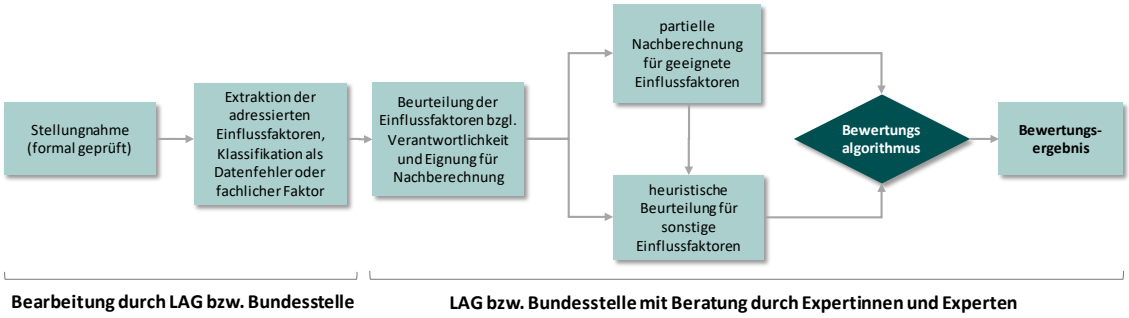

Abbildung 32: Ablauf der fachlichen Bewertung von Stellungnahmen im Überblick

Nach Abschluss der formalen Prüfung (siehe Abschnitt 6.5) werden anhand der in der jeweiligen Stellungnahme angeführten Gründe die adressierten Einflussfaktoren von der LAG bzw. Bundesstelle extrahiert[64] und ihre Art entweder als Datenfehler oder als fachlich-inhaltlicher Einflussfaktor klassifiziert. Im nächsten Schritt wird jeder der Datenfehler oder fachlich-inhaltlichen Einflussfaktoren dahingehend beurteilt, ob es sich um einen vom Leistungserbringer zu verantwortenden Faktor handelt, und ob sein Beitrag zum Indikatorergebnis so eindeutig ist, dass er bei der Indikatorberechnung berücksichtigt werden könnte (siehe Abschnitt 6.4.1).

Im darauffolgenden Schritt wird zum einen auf Grundlage der dafür als geeignet eingeschätzten Einflussfaktoren eine Nachberechnung des Indikatorwertes für den stellungnehmenden Leistungserbringer vorgenommen. Diese wird im folgenden als partielle Nachberechnung bezeichnet, da sie ggf. nicht alle Einflussfaktoren berücksichtigen kann. Zum anderen werden die weiteren, nicht für eine Nachberechnung geeigneten Einflussfaktoren mit Unterstützung der Expertinnen und Experten – und unter Berücksichtigung der Ergebnisse der partiellen Nachberechnung – heuristisch auf Grundlage fachlicher Erfahrung beurteilt. Mittels eines Bewertungsalgorithmus, der der unterschiedlichen Art der in den Stellungnahmen geltend gemachten Einflussfaktoren Rechnung trägt, werden die Ergebnisse aus partieller

Leistungserbringer mit Stellungnahme mit originalen Indikatorergebnissen der Leistungserbringer ohne Stellungnahme sind methodisch nicht angemessen.

[64] Damit ist nicht gemeint, dass jeder klinische Parameter, der möglicherweise in einer Stellungnahme erläutert wird, separat aufgeführt werden soll, sondern nur der Einflussfaktor oder die Konstellation von Einflussfaktoren, die vom Leistungserbringer als besonders angesehen wird (z. B. „Vorliegen einer Kontraindikation“ oder „Multimorbidität“).

Nachberechnung und heuristischer Beurteilung anschließend zu einem Bewertungsergebnis zusammengeführt.

Die Klassifikation der aus den Stellungnahmen extrahierten Einflussfaktoren hinsichtlich ihrer Eignung für eine Nachberechnung und insbesondere hinsichtlich der Verantwortlichkeit des Leistungserbringers erfordert fachliche Expertise. Das gleiche gilt für die Beurteilung der nicht für eine Nachberechnung geeigneten Einflussfaktoren. Die LAG bzw. Bundestelle wird deshalb von Expertinnen und Experten in Form einer Fachkommission beraten (siehe Abschnitt 6.6). Dadurch soll eine angemessene Berücksichtigung der oft komplexen Sachverhalte in den Stellungnahmen gewährleistet werden. Die partielle Nachberechnung erfolgt also zwangsläufig im Austausch mit der Fachkommission, was eine angemessene technische Unterstützung erfordert (siehe Abschnitt 6.8). Dabei sollen die Leistungserbringer gegenüber den Mitgliedern der Fachkommission pseudonymisiert werden, um Verzerrungen der fachlichen Beurteilung zu vermeiden.

Die LAG bzw. Bundesstelle trifft die abschließende Entscheidung in Bezug auf die Einordnung des Indikatorergebnisses in das Bewertungsschema. Weicht die LAG bzw. Bundesstelle von den Empfehlungen der Fachkommission ab, begründet und dokumentiert sie diese Abweichung. Mit der abschließenden Entscheidung in Bezug auf die Einordnung des Indikatorergebnisses in das Bewertungsschema ist die Qualitätsbewertung abgeschlossen.

6.4 Bewertungsalgorithmus

In den folgenden Abschnitten werden die Details des empfohlenen Vorgehens bei der Bewertung von Stellungnahmen und der partiellen Nachberechnung von Indikatorwerten erläutert. Abbildung 33 zeigt den grundlegenden Bewertungsalgorithmus im Überblick. Ziel des Algorithmus ist, zu einer möglichst objektiven und einheitlichen Einstufung in die Kategorien des in Abschnitt 6.1 entwickelten Bewertungsschemas zu kommen (siehe Tabelle 14). Er ist daher so angelegt, dass die Beurteilungsschritte so weit wie möglich auf expliziten, quantitativen Regeln basieren. Dies ist insbesondere für die unter dem Gesichtspunkt der *accountability* maßgeblichen Ergebnisse auf erster Ebene des Bewertungsschemas, „Qualitätsdefizit“ versus „Hinweis auf Qualitätsdefizit entkräftet“, von Bedeutung. Verbleibende Unsicherheit wird in nachfolgenden, qualitativen Beurteilungsschritten adressiert.

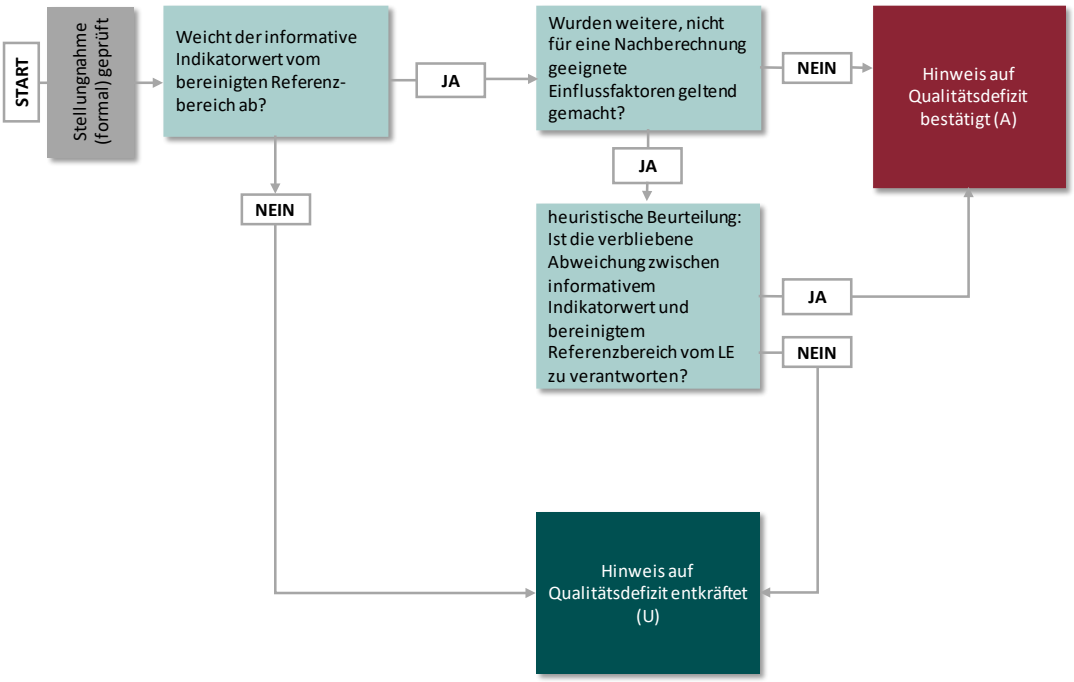

Abbildung 33: Grundlegender Algorithmus der fachlichen Bewertung

Im ersten Schritt wird eine (partielle) Nachberechnung des Indikatorwertes für den Leistungserbringer und ein Vergleich mit einem entsprechend modifizierten Referenzbereich vorgenommen. Dazu wird ein Indikatorwert berechnet, der um diejenigen nicht vom Leistungserbringer zu verantwortenden Datenfehler und fachlich-inhaltlichen Einflussfaktoren bereinigt ist, deren Auswirkung auf das Indikatorergebnis eindeutig benannt werden kann. Diese Bereinigung erfolgt z. B. dadurch, dass Behandlungsfälle, die vom Indikator nicht hätten adressiert werden sollen, aus der Grundgesamtheit des Indikators ausgeschlossen werden (Details siehe Abschnitt 6.4.1). Der so bereinigte Wert wird im Folgenden als *informativer Indikatorwert* bezeichnet und dient ausschließlich als Information im Rahmen der fachlichen Bewertung; er hat nicht die Funktion, den originalen Indikatorwert[65] der Einrichtung zu ersetzen.

Desweiteren wird aus dem originalen Referenzbereich[66] des Indikators ein modizifierter Referenzbereich abgeleitet. Dieser ist – wie der informative Indikatorwert – um den Anteil bereinigt, der schon im originalen Referenzbereich für die gerichteten, nicht vom Leistungserbringer zu verantwortenden Einflussfaktoren einkalkuliert ist (siehe Abschnitt 6.4.2). Dieser Referenzbereich kann als der Referenzbereich verstanden werden, der festgelegt worden wäre, wenn die Berechnungsvorschrift des bereinigten Indikators von vorneherein die gerichteten Einflüsse berücksichtigt hätte, und wird im Folgenden als *bereinigter Referenzbereich* bezeichnet.

[65] Unter „originaler Indikatorwert“ wird in diesem Bericht der Indikatorwert verstanden, der anhand der gültigen Rechenregeln des Indikators im Regelbetriebs berechnet und in den Auswertungen der Leistungserbringer dargestellt wird.
[66] Mit „originaler Referenzbeich“ wird in diesem Bericht der vom G-BA beschlossene Referenzbereich eines Indikators bezeichnet, der bei der Erstellung der Auswertungen im Regelbetrieb angewendet wird.

Der informative Indikatorwert wird mit dem bereinigten Referenzbereich unter Anwendung der üblichen Vergleichsregel (statistisches Verfahren) verglichen. Wenn bei diesem Vergleich der informative Indikatorwert den bereinigten Referenzbereich nicht verfehlt, ist die ursprüngliche Abweichung (des originalen Indikatorwerts vom originalen Referenzbereich) hinreichend durch die besondere Konstellation erklärt, die vom Leistungserbringer geltend gemacht wurde. In diesem Fall lautet die Bewertung des Indikatorergebnisses „Hinweis auf Qualitätsdefizit entkräftet". Soll genauer gekennzeichnet werden, ob Datenfehler, fachlich-inhaltliche Einflussfaktoren oder eine Kombination aus beiden Arten von Faktoren für das ursprüngliche Verfehlen des Referenzbereichs ursächlich waren, kann anschließend eine Binnendifferenzierung dieser Unterkategorien auf Grundlage weiterer Berechnungen erfolgen (siehe Abschnitt 6.4.5).

Erreicht dagegen der informative Indikatorwert den bereinigten Referenzbereich nicht, ist zu prüfen, ob vom Leistungserbringer zusätzliche Einflüsse geltend gemacht wurden, die in der partiellen Nachberechnung nicht berücksichtigt werden konnten. Ist dies nicht der Fall, liegen in der Stellungnahme keine Belege vor, die das ursprüngliche Indikatorergebnis entkräften. Die Qualitätsanforderung des Indikators gilt damit als nicht erfüllt und die Einstufung des Indikatorergebnisses lautet „Qualitätsdefizit".

Wurden dagegen in der Stellungnahme des Leistungserbringers über die in der Nachberechnung berücksichtigten Faktoren hinaus zusätzliche Einflüsse auf den Indikatorwert angeführt, werden diese durch die Expertinnen und Experten der Fachkommission qualitativ beurteilt. Bei dieser Beurteilung wird – ausgehend vom informativen Indikatorwert und vom bereinigten Referenzbereich – eingeschätzt, ob das Ausmaß dieser verbliebenen, zusätzlichen Einflüsse das Verfehlen des bereinigten Referenzbereiches trotz partieller Nachberechnung erklärt. Sofern Datenfehler nicht vom Leistungserbringer zu verantworten sind und nicht bereits bei der partiellen Nachberechnung berücksichtigt werden konnten, sollen sie ebenfalls in diese Bewertung mit einbezogen werden. Die zu beurteilende Frage lautet:[67] „*Ist es schlüssig, dass die verbliebene Abweichung des Indikatorwerts des Leistungserbringers vom Referenzbereich nach partieller Nachberechnung vollständig durch die restlichen, nicht vom Leistungserbringer zu verantwortenden Faktoren und Datenfehler erklärt ist?*". In Abhängigkeit von der Beantwortung dieser Frage erfolgt dann die Einstufung als „Hinweis auf Qualitätsdefizit entkräftet" (wenn die Abweichung schlüssig nicht vom Leistungserbringer zu verantworten ist) bzw. als „Qualitätsdefizit".

Um bei einer Einstufung als „Qualitätsdefizit" genauer zwischen Datenfehlern und einem Defizit der Versorgungsqualität zu differenzieren, kann anschließend eine Binnendifferenzierung dieser Unterkategorien erfolgen (siehe Abschnitt 6.4.4).

Die Details dieses Bewertungsalgorithmus werden in den folgenden Abschnitten ausführlich erläutert.

[67] Siehe Abschnitt 6.4.3 für methodische Limitationen bei verteilungsbezogenen Referenzbereichen.

6.4.1 Analyse der in den Stellungnahmen genannten Gründe

Bevor der informative Indikatorwert im ersten Beurteilungsschritt berechnet werden kann, muss zunächst anhand der in einer Stellungnahme angeführten Gründe für das Verfehlen des Referenzbereichs analysiert werden, welche Einflussfaktoren auf das Indikatorergebnis vom Leistungserbringer geltend gemacht werden. Beispielsweise können sich die angeführten Gründe auf ein besonderes Risikoprofil der Patientinnen und Patienten, auf Notfallsituationen oder auf Datenfehler beziehen. Für jeden dieser Einflussfaktoren sind folgende Fragen zu klären:

- Handelt es sich bei dem Einflussfaktor um einen Fehler in den QS-Daten (Datenfehler) oder um einen fachlich-inhaltlichen Einflussfaktor (besondere Versorgungskonstellation)?
- Ist der Datenfehler oder der fachlich-inhaltliche Einflussfaktor vom Leistungserbringer zu verantworten oder hatte der Leistungserbringer keine Möglichkeit, diesen Faktor zu beeinflussen?
- Ist die Auswirkung des Datenfehlers oder des fachlich-inhaltlichen Einflussfaktors auf das Indikatorergebnis so eindeutig zu beziffern, dass dieser Einflussfaktor bei einer Nachberechnung berücksichtigt werden kann?

Datenfehler vs. fachlich-inhaltlicher Einflussfaktor

Neben fachlich-inhaltlichen Einflussfaktoren können auch Fehler in den übermittelten Daten dazu führen, dass ein Indikator die Versorgungsqualität eines Leistungserbringers nicht angemessen abbildet. Beispielsweise können Behandlungsfälle in die Indikatorberechnung für einen Leistungserbringer eingegangen sein, die bei korrekten Daten nicht hätten berücksichtigt werden dürfen. Um eine nachvollziehbare Einstufung des Indikatorergebnisses zu erreichen, muss daher die Beurteilung von in den Stellungnahmen angeführten Datenfehlern mit der Beurteilung angeführter fachlich-inhaltlicher Einflussfaktoren verknüpft werden. Die Klassifizierung, bei welchen der in der Stellungnahme eines Leistungserbringers angeführten Einflussfaktoren es sich um Datenfehler handelt, nimmt die LAG bzw. Bundesstelle vor.

Die fachliche Bewertung bezieht sich prinzipiell auf das Indikatorergebnis einer Einrichtung, das auf Grundlage der regulär dokumentierten Fälle berechnet wurde. Eine Nachdokumentation von Behandlungsfällen, die fälschlich nicht dokumentiert wurden, ist im Rahmen des Stellungnahmeverfahrens nicht vorgesehen. Solche zusätzlichen Behandlungsfälle können bei der Bewertung der Stellungnahmen nicht berücksichtigt und nicht als besondere Konstellation anerkannt werden, da sich ihre Auswirkung auf den Indikatorwert einer Einrichtung nicht angemessen beurteilen lässt.

Beurteilung der Verantwortlichkeit für einen Einflussfaktor

Sowohl bei Datenfehlern als auch bei fachlich-inhaltlichen Einflussfaktoren ist zwischen solchen Faktoren zu unterscheiden, die vom Leistungserbringer nicht zu verantworten sind, und solchen, die der Leistungserbringer zu verantworten hat (Beispiele siehe Anhang, Kapitel 5). Die Beurteilung erfolgt auf Basis der in den Stellungnahmen vorgebrachten Gründe und dazugehörigen Belege.

Die Beurteilung, welche der in der Stellungnahme eines Leistungserbringers angeführten Datenfehler vom Leistungserbringer zu verantworten sind, trifft die LAG bzw. Bundesstelle. Die Beurteilung, ob ein fachlich-inhaltlicher Einflussfaktor vom Leistungserbringer zu verantworten ist, erfolgt grundsätzlich unter Beratung der LAG bzw. Bundesstelle durch die Expertinnen und Experten der Fachkommission. In diese Beurteilung soll auch einfließen, ob vom Leistungserbringer mit der gebotenen Sorgfaltspflicht alles Zumutbare getan wurde, um den aus seiner Sicht nicht von ihm zu verantwortenden Einflussfaktor aufzufangen (z. B. durch die Einplanung von genügend Ressourcen und genügend Zeit).

Bei manchen Einflussfaktoren kann es erforderlich sein, die Verantwortlichkeit des Leistungserbringers je nach Konstellation des Behandlungsfalls unterschiedlich einzuschätzen. Beispielsweise könnte eine Wiederbelebungsmaßnahme bei einem Patienten oder einer Patientin als vom Leistungserbringer nicht zu verantwortender Faktor eingeschätzt werden, wenn sie vor Eintreffen in der Einrichtung erfolgt ist, aber nicht, wenn sie erst im weiteren Behandlungsverlauf notwendig wurde. Für solche Einflussfaktoren muss entschieden werden, ob die Verantwortlichkeit für jeden Behandlungsfall eindeutig als gegeben oder nicht gegeben klassifiziert werden kann und sie sich für eine partielle Nachberechnung auf Grundlage der anerkannten Behandlungsfälle eignen (siehe unten), oder ob die Auswirkung auf das Indikatorergebnis für diesen Einflussfaktor nicht zuverlässig angegeben werden kann.

Auswirkung auf das Indikatorergebnis

Ein wichtiges Merkmal der in den Stellungnahmen adressierten Einflussfaktoren ist das Ausmaß, in dem sie das Indikatorergebnis beeinflussen. Damit ist gemeint, dass sich für manche Einflussfaktoren zuverlässig angeben lässt, wie sie bei einer Nachberechnung des Indikatorergebnisses berücksichtigt werden sollten, während dies für andere Einflussfaktoren nicht möglich ist. Wie groß die Auswirkung eines geltend gemachten Einflussfaktors auf das Indikatorergebnis ist, hängt unmittelbar von seiner Beeinflussbarkeit durch den Leistungserbringer ab (siehe IQTIG 2019a). Die Beurteilung, ob ein Einflussfaktor durch partielle Nachberechnung berücksichtigt werden kann, erfolgt daher zusammen mit der Beurteilung der Verantwortlichkeit durch die LAG bzw. Bundesstelle unter Beratung durch die Expertinnen und Experten der Fachkommission.

Beschreibt ein in der Stellungnahme angeführter Grund einen Einflussfaktor, bei dessen Vorliegen die betreffenden Behandlungsfälle nicht vom Qualitätsindikator hätten adressiert werden sollen, d. h. nicht zur Zielpopulation gehören, so kann das Ausmaß auf den Indikatorwert eindeutig angegeben werden. Beispielsweise soll ein Indikator zur Sterblichkeit durch Pneumonie die Behandlungskompetenz der Leistungserbringer in Bezug auf die Verhinderung krankheitsbedingter Todesfälle messen. Das Qualitätsmerkmal bezieht sich auf Patientinnen und Patienten, die mit dem Ziel der Heilung behandelt werden und nicht auf solche, die mit palliativer Intention behandelt werden[68]. Für den Einflussfaktor palliative Behandlungsintention lässt sich

[68] Im Leistungsbereich „Ambulant erworbene Pneumonie“ wird zwar über ein Datenfeld erhoben, ob mit palliativer Intention behandelt wird. Derzeit wird dieses Datenfeld jedoch nicht als valide genug erachtet, um Fälle mit Therapieverzicht aus der Grundgesamtheit des Indikators zur Sterblichkeit auszuschließen. In diesem Fall erweist es sich schwierig, die Fälle, die man ausschließen möchte, durch ein Datenfeld zu identifizieren. Dieser Einflussfaktor muss daher im Stellungnahmeverfahren berücksichtigt werden.

bei diesem Indikator also klar entscheiden, wie er bei der Indikatorberechnung berücksichtigt werden sollte. Die partielle Nachberechnung erfolgt bei solchen Einflussfaktoren dadurch, dass die Behandlungsfälle, bei denen das Vorliegen des Einflussfaktors anerkannt wird, aus der Grundgesamtheit (Nenner) des Indikators ausgeschlossen werden (siehe Anhang, Kapitel 6). Bei den für eine Nachberechnung geeigneten Einflussfaktoren handelt es sich oft um besondere Patientencharakteristika wie Kontraindikationen für eine Behandlung oder das Auftreten unerwarteter Notfallsituationen. Auch andere Einflussfaktoren, bei denen keine Verantwortlichkeit des Leistungserbringers für diesen Faktor anzunehmen ist (z. B. ein extern bedingter Stromausfall), eignen sich für eine Berücksichtigung im Rahmen der partiellen Nachberechnung.

Andere Einflussfaktoren erhöhen hingegen lediglich die Wahrscheinlichkeit für ein ungünstigeres Indikatorergebnis. Ein Beispiel hierfür sind Begleiterkrankungen der Patientinnen und Patienten, die bisher in der Risikoadjustierung des Indikators nicht oder nicht adäquat berücksichtigt ist. So könnte etwa das Vorliegen von Übergewicht durch die Begünstigung lokaler Komplikationen bei Operationswunden das Ergebnis eines Qualitätsindikators beeinflussen, der Komplikationshäufigkeiten einer Behandlung misst. Werden Informationen zu solchen Einflussfaktoren nicht erfasst und im Risikoadjustierungsmodell des Indikators nicht berücksichtigt, lässt sich ihr Einfluss auf den Indikatorwert eines Leistungserbringers nicht eindeutig benennen. Kommen solche Faktoren bei einem Leistungserbringer häufiger vor als in der Gesamtpopulation der Leistungserbringer, so können sie bei der fachlichen Bewertung nicht im Rahmen der partiellen Nachberechnung, sondern nur im qualitativen Beurteilungsschritt berücksichtigt werden.

Je nach betroffenem Datenfeld können auch Datenfehler unterschiedlich gut durch eine Nachberechnung adressiert werden. Wurden durch Datenfehler Behandlungsfälle fälschlich in die Indikatorberechnung einbezogen, so lässt sich deren Auswirkung durch Ausschluss der Fälle aus der Grundgesamtheit des Indikators ermitteln. Hat ein Datenfehler dazu geführt, dass das interessierende Ereignis (z. B. Komplikationen oder die Durchführung eines Prozesses) fälschlich als gegeben bzw. als nicht gegeben gewertet wurde, so kann das Ausmaß des Datenfehlers auf den Indikatorwert ebenfalls einfach angegeben werden. Die Nachberechnung besteht dann darin, den Behandlungsfall als Behandlungsfall ohne interessierendes Ereignis zu werten, also ihn z. B. aus dem Zähler des Indikators auszuschließen (siehe Anhang, Kapitel 6). Für Datenfehler muss also entschieden werden, ob sie durch Ausschluss der Behandlungsfälle aus der Grundgesamtheit oder nur aus dem Zähler korrigiert werden können. Geht ein fehlerhafter Wert dagegen eher indirekt in das Indikatorergebnis ein, beispielsweise wenn Werte für Risikofaktoren in einer falschen Einheit dokumentiert wurden, so könnte die Auswirkung des Datenfehlers auf den Indikatorwert nur mit hohem Aufwand ermittelt werden.

Weitere Beispiele für die Beurteilung von Einflussfaktoren sind im Anhang, Kapitel 5 dargestellt. Eine abschließende Auflistung und Kategorisierung aller denkbaren Datenfehler und fachlich-inhaltlichen Einflussfaktoren mit Angabe von Verantwortlichkeit des Leistungserbringers und Eignung für eine partielle Nachberechnung ist nicht möglich, da immer wieder neue besondere Konstellationen auftreten können. Die Beurteilungsergebnisse von in früheren Stellungnahmen angegebenen Gründen können daher zwar als Orientierung genutzt werden, um eine

Einheitlichkeit der Beurteilung zu fördern, die abschließende Beurteilung der Gründe und Einflussfaktoren muss aber jeweils auf Grundlage der jeweils aktuellen Stellungnahme erfolgen.

Der informative Indikatorwert, der die Grundlage für den quantitativen Teil der fachlichen Bewertung ist, wird ausgehend vom originalen Indikatorwert berechnet, indem die auszuschließenden Behandlungsfälle und die auszuschließenden interessierenden Ereignisse berücksichtigt werden. Details zur Berechnung sind in Anhang, Kapitel 6 dargestellt. Alle anderen Einflussfaktoren, die sich nicht diesen beiden Kategorien zuordnen lassen, gehen nicht in die partielle Nachberechnung ein, sondern werden durch die LAG bzw. Bundesstelle und die Expertinnen und Experten der Fachkommission heuristisch beurteilt.

6.4.2 Festlegung von Referenzbereichen für die partielle Nachberechnung

Für die quantitative Beurteilung des Indikatorwerts eines Leistungserbringers, der um gerichtete, nicht vom Leistungserbringer zu verantwortende Einflüsse bereinigt ist, wird ein modifizierter Referenzbereich benötigt. Dies ist deshalb erforderlich, weil die Referenzbereiche der externen Qualitätssicherung eine Abweichung vom Idealwert erlauben, die aus zwei Anteilen besteht:

1. Zum einen sind gerichtete Einflüsse auf den Indikatorwert einkalkuliert, die nicht erfasst werden und die nicht vom Leistungserbringer zu verantworten sind (siehe Abschnitt 3.2.1). Beispielsweise könnte ein Ergebnisindikator die Komplikationshäufigkeit nach einer Operation messen, die im Idealfall 0 % betrüge. Aufgrund systematischer, nicht erfasster Einflüsse wie pathophysiologischer Gesetzmäßigkeiten oder der Charakteristika der notwendigen Operationen sei angenommen, dass bei dem betreffenden Krankheitsbild und Operationsverfahren eine mittlere Komplikationshäufigkeit von 5 % selbst bei bester Versorgungsqualität zu erwarten ist. Würden die zugrunde liegenden Indikatorwerte nur um den Betrag dieser Einflussfaktoren (im Beispiel 5 %) vom Idealwert abweichen, läge optimale Qualität im Sinn des Qualitätsziels des jeweiligen Indikators vor.
2. Zum anderen berücksichtigen die Referenzbereiche, dass nicht jeder Leistungserbringer optimale Qualität erreichen kann. Das Erreichen des Referenzbereichs repräsentiert also nicht optimale Qualität, sondern die Erfüllung erwartbarer Standards (zureichende Qualität), die von jedem Leistungserbringer verlangt werden kann. Es wird damit anerkannt, dass es oberhalb der Schwelle zu zureichender Qualität weitere Qualitätsunterschiede zwischen den Leistungserbringern geben kann (z. B. besonders gute Erfüllung der Anforderungen). Würden im obigen Indikatorbeispiel 2 Prozentpunkte zusätzliche Abweichung toleriert, läge die Referenzbereichsgrenze bei 5 % + 2 % = 7 %.

Für die Bewertung nach partieller Nachberechnung sind diese beiden Anteile des Referenzbereichs von großer Bedeutung: Es wird angenommen, dass durch die Stellungnahme gerichtete Einflussfaktoren, die zum Zeitpunkt der Indikatorberechnung noch nicht berücksichtigt waren, bekannt werden. Wird nun eine partielle Nachberechnung vorgenommen und würde der informative Indikatorwert mit dem originalen Referenzbereich verglichen werden, würden die gerichteten Einflussfaktoren auf den Indikatorwert doppelt berücksichtigt werden: Einmal im Referenzbereich, bei dessen Setzung sie einkalkuliert wurden, und einmal

durch den Vorgang der partiellen Nachberechnung des Indikatorwerts. Für einen angemessenen Vergleich mit dem informativen Indikatorwert muss der Referenzbereich also um den Anteil gerichteter Einflussfaktoren verringert werden (siehe Abbildung 34). Der Anteil des Referenzbereiches, der eine Toleranz für zulässige Qualitätsunterschiede ausdrückt (Nr. 2), bleibt dagegen auch nach partieller Nachberechnung erhalten.[69]

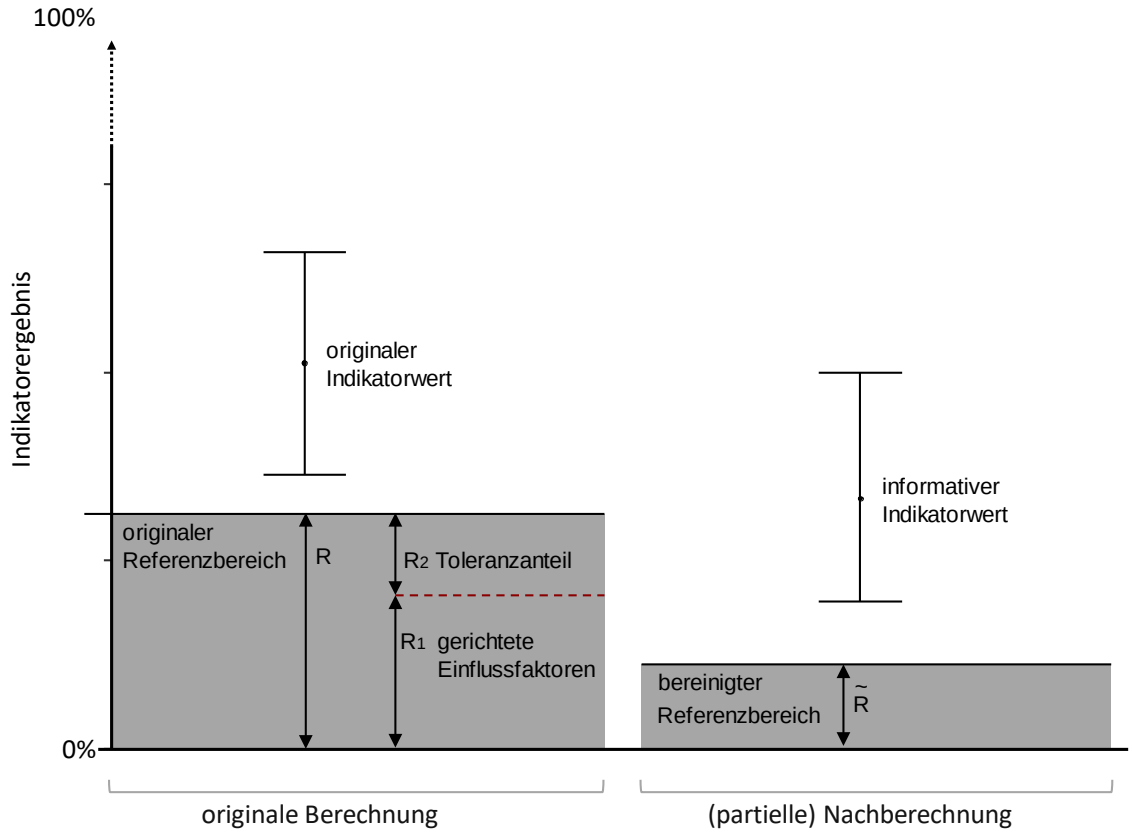

Abbildung 34: Beispiel für die Nachberechnung von Indikatorergebnis und Referenzbereich zur fachlichen Bewertung. Links: Das ursprüngliche Indikatorergebnis vor der fachlichen Bewertung. Rechts: Nachberechnung von Referenzbereich und Indikatorergebnis.

Der so „bereinigte" Referenzbereich kann als der Referenzbereich verstanden werden, der festgelegt worden wäre, hätte die Rechenregel des Indikators von vorneherein die gerichteten Einflüsse berücksichtigt. Beim Vergleich von informativem Indikatorwert mit dem bereinigten Referenzbereich gibt der Referenzbereich also an, welche vom Leistungserbringer zu verantwortende Abweichung vom Idealwert des Indikators (entweder 0 % oder 100 %) noch toleriert wird, bevor ein Qualitätsdefizit attestiert wird (im obigen Beispiel eine Abweichung von ca. 2 %). Weitere Details zur empfohlenen Berechnungsvorschrift sind in Anhang, Kapitel 6 dargestellt.

[69] Für verteilungsbasierte Referenzbereiche (z. B. perzentilbasierte) gelten bestimmte Einschränkungen, die in Abschnitt 6.4.3 näher erläutert sind.

Zum Zeitpunkt der Berichtslegung entwickelt das IQTIG eine Methodik, wie bei Festlegung von Referenzbereichen die Größe der beiden Anteile des Referenzbereichs angemessen und nachvollziehbar festgelegt werden kann. Für die bisher in der externen Qualitätssicherung eingesetzten Referenzbereiche wurde jedoch die jeweilige Größe der beiden Anteile nicht gesondert ausgewiesen. Sie muss daher zunächst retrospektiv geschätzt werden, bevor auf dieser Basis ein bereinigter Referenzbereich berechnet werden kann. Ein mögliches Vorgehen für diese retrospektive Abschätzung ist ebenfalls in Anhang, Kapitel 6 dargestellt.

6.4.3 Methodische Limitationen verteilungsbezogener Referenzbereiche

Verteilungsbezogene Referenzbereiche erlauben keine Aussagen über die Erfüllung definierter Standards, sondern lediglich vergleichende Aussagen relativ zu den Ergebnissen aller Leistungserbringer (IQTIG 2019a: 161 f.). Ob beispielsweise ein Leistungserbringer, der im 5. Perzentil der Verteilung aller Leistungserbringer in diesem Qualitätsindikator liegt, erwartbare Standards in der Versorgung erfüllt, kann nicht aus dem Indikatorergebnis erschlossen werden. Durch ein Stellungnahmeverfahren kann zwar die Validität der Messung des Indikators für einen Leistungserbringer geprüft werden, jedoch keine umfassende Qualitätsaussage zu einem Leistungserbringer (siehe Abschnitt 3.3 für eine ausführliche Begründung) getroffen werden. Daher kann die fachliche Bewertung bei verteilungsbezogenen Referenzbereichen zwar ergeben, dass der Vergleich des Indikatorwerts eines Leistungserbringers mit dem Referenzbereich nicht angemessen ist, aber nicht, ob der Leistungserbringer mit diesem Indikatorwert einen definierten Standard erreicht. Denn wenn ein Standard für die Versorgung nicht schon bei der Entwicklung des Indikators festgelegt werden konnte, kann dies nicht im Rahmen eines Stellungnahmeverfahrens geschehen.

Beispielsweise könnte sich im Stellungnahmeverfahren mit einem Leistungserbringer bestätigen, dass keine besonderen Konstellationen Einfluss auf den Indikatorwert genommen haben und dieser den Erfüllungsgrad des Qualitätsziels des Indikators angemessen darstellt. Dennoch könnte das Ergebnis des Leistungserbringers absolut gesehen keinen Qualitätsmangel darstellen, sondern lediglich schlechter sein als das Ergebnis von z. B. 90 % der Leistungserbringer (bei einem Referenzbereich auf Basis des 10. Perzentils).

Strenggenommen müsste daher aus methodischer Sicht für Indikatoren mit einem verteilungsbezogenen Referenzbereich das Einstufungsergebnis „keine Qualitätsbewertung“ (S) lauten.[70] Da derzeit für viele Qualitätsindikatoren der externen Qualitätssicherung noch keine festen Referenzbereiche definiert sind, hätte ein solches Vorgehen den Nachteil, dass trotz aufwändiger Datenerhebung keine Qualitätsaussage getroffen werden würde. Als Interimslösung empfiehlt das IQTIG daher, perzentilbasierte Referenzbereiche im

[70] Denkbar wäre grundsätzlich auch, in die Klassifikation des Bewertungsergebnisses weitere Unterkategorien für verteilungsbezogene Referenzbereiche einzufügen. Diese könnten dann deutlich machen, dass z. B. „Qualitätsdefizit“ in diesem Fall keinen definierten Qualitätsmangel darstellt, sondern „schlechter als andere Leistungserbringer“ bedeutet. Von einer solchen Aufteilung in zusätzliche Kategorien wird aber abgeraten, da sie für die Nutzung der Informationen im Rahmen von *accountability* (z. B. für die Qualitätsberichte der Krankenhäuser oder für Vergütungsentscheidungen; siehe Abschnitt 2.2) weniger verständlich sein dürfte, den Blick auf wichtige Informationen erschwert und zudem keine Konsequenzen für Handlungsentscheidungen hat.

Stellungnahmeverfahren wie feste Referenzbereiche zu behandeln. Perspektivisch sollte durch Festlegung fester, kriteriumsbezogener Referenzbereiche für diese Indikatoren Klarheit geschaffen werden, welche Qualitätsstandards von den Leistungserbringern gefordert werden.

6.4.4 Vergabe der Unterkategorien für die Einstufung „Qualitätsdefizit“

Wurde der indikatorbasierte Hinweis auf ein Defizit der Versorgungsqualität nach Berücksichtigung aller vom Leistungserbringer nicht zu verantwortenden Einflussfaktoren nicht entkräftet, wird als Ergebnis der fachlichen Bewertung ein „Qualitätsdefizit“ festgestellt. Das Bewertungsschema zum Stellungnahmeverfahren sieht Unterkategorien (Ziffern) vor, die unterscheiden, ob Datenfehler oder fachlich-inhaltliche Faktoren oder eine Kombination aus beiden entscheidend für die Bewertung als „Qualitätsdefizit“ waren (siehe Abschnitt 6.1). Im Folgenden wird ein Vorgehen für einheitliche und nachvollziehbare Kategorisierungen in die Unterkategorien für ein Qualitätsdefizit beschrieben. Das Vorgehen dazu wird in Abbildung 35 visualisiert.

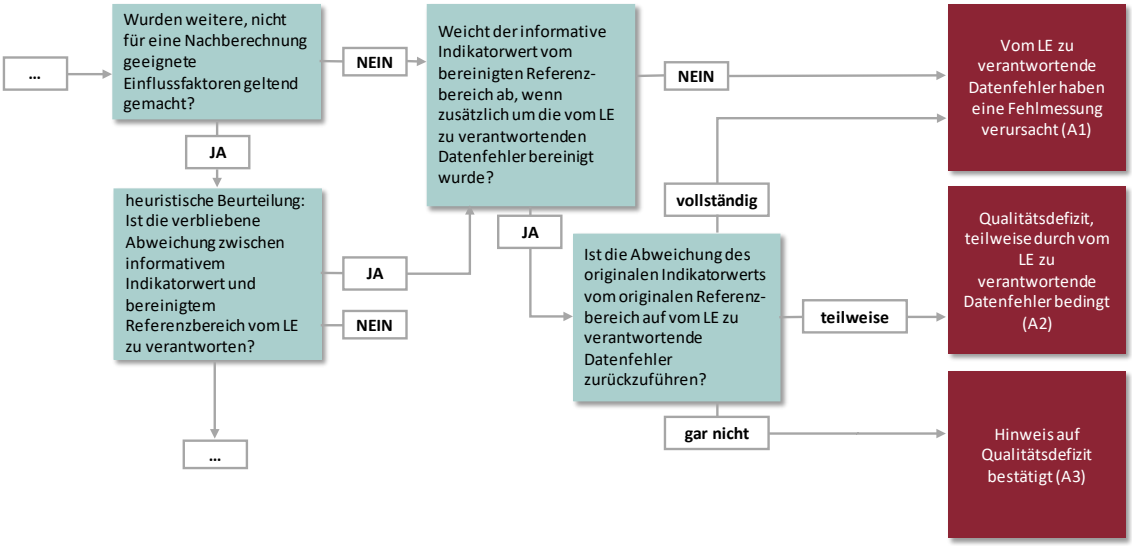

Abbildung 35: Algorithmus zur Differenzierung der Gründe für die Einstufung als Qualitätsdefizit

Der erste Schritt der Differenzierung besteht in einer expliziten Beurteilungsregel, um eine einheitliche Beurteilung sicherzustellen. Dazu wird eine Variante des informativen Indikatorwerts berechnet, bei der nicht nur Datenfehler berücksichtigt werden, die nicht vom Leistungserbringer zu verantworten sind, sondern alle Datenfehler, soweit das rechnerisch eindeutig möglich ist. Diese zusätzliche partielle Nachberechnung erfolgt analog zu dem in den vorherigen Abschnitten geschilderten Vorgehen. Führt diese modifizierte Berechnung des informativen Indikatorwerts dazu, dass der bereinigte Referenzbereich nicht mehr verfehlt wird, ist das ursprüngliche Indikatorergebnis vollständig durch die Kombination aus nicht vom Leistungserbringer zu verantwortenden Einflussfaktoren (falls vorhanden) sowie von ihm zu verantwortenden Datenfehlern erklärt. Das Einstufungsergebnis lautet in diesem Fall „Vom Leistungserbringer zu verantwortende Datenfehler haben eine Fehlmessung verursacht“ (A1). An den entsprechenden Leistungserbringer werden damit hinsichtlich der Beurteilung, inwieweit

(zusätzlich) ein Defizit der Versorgungsqualität vorgelegen hatte, die gleichen Anforderungen gestellt wie an andere Leistungserbringer ohne Datenfehler.

Erklärt die zusätzliche Berücksichtigung der vom Leistungserbringer zu verantwortenden Datenfehler das Qualitätsdefizit dagegen nicht, so ist in der Regel zumindest anteilig von Defiziten der Versorgungsqualität auszugehen. Die Einstufung in die passende erläuternde Unterkategorie A2 oder A3 soll in diesem Fall von der LAG bzw. Bundesstelle mit Unterstützung durch die beratenden Expertinnen und Experten der Fachkommission vorgenommen werden. Sie soll sich daran orientieren, in welchem Ausmaß die Datenfehler einerseits und die inhaltlichen Einflussfaktoren andererseits ausschlaggebend für die Bewertung als „Qualitätsdefizit“ waren. Eine exakte Quantifizierung des jeweiligen Anteils ist jedoch nicht möglich, da Datenfehler zum Verfehlen des Referenzbereiches beigetragen haben können, deren Auswirkung auf den Indikatorwert sich nicht über eine Nachberechnung exakt angeben lässt, und da für einzelne Behandlungsfälle, die in die Indikatorberechnung eingegangen sind, gleichzeitig Datenfehler und fachlich-inhaltliche Faktoren relevant sein können.

In seltenen Fällen könnte das ursprüngliche Abweichen vom Referenzbereich vollständig durch vom Leistungserbringer zu verantwortende Datenfehler verursacht sein, die nicht durch eine Nachberechnung adressiert werden können (denkbar wären beispielsweise gehäufte Eingabefehler bei einem für die Risikoadjustierung genutzten Datenfeld). Für solche Konstellationen sieht der Algorithmus die Möglichkeit vor, auf Grundlage der nicht-quantativen Einschätzung die Unterkategorie A1 zu vergeben.

6.4.5 Vergabe der Unterkategorien bei Entkräftung des Hinweises auf ein Qualitätsdefizit

Falls auch Transparenz über die Gründe gewünscht wird, die dazu geführt haben, dass der indikatorbasierte Hinweis auf ein Qualitätsdefizit entkräftet wurde, können die entsprechenden Unterkategorien des Bewertungsschemas genutzt werden, die zwischen Datenfehlern und fachlich-inhaltlichen Faktoren unterscheiden (U1 bis U3; siehe Abschnitt 6.1). Wenn eine zuverlässige Differenzierung angestrebt wird, sollte diese entsprechend den in Kapitel 5 erläuterten methodischen Anforderungen möglichst anhand expliziter, quantitativer Kriterien erfolgen.

Gegenüber dem eingangs dargestellten grundlegenden Algorithmus (siehe Abbildung 33) wird der Einstufungsprozess nach Entkräftung des Hinweises auf ein Qualitätsdefizit an zwei Stellen weitergeführt, die im Folgenden beschrieben werden.

Differenzierung nach Entkräftung durch partielle Nachberechnung

Um eine Differenzierung vorzunehmen, wenn die partielle Nachberechnung des Indikatorergebnisses zur Entkräftung des Hinweises auf ein Qualitätsdefizit geführt hat, sollte der Bewertungsalgorithmus wie in Abbildung 36 dargestellt erweitert werden.

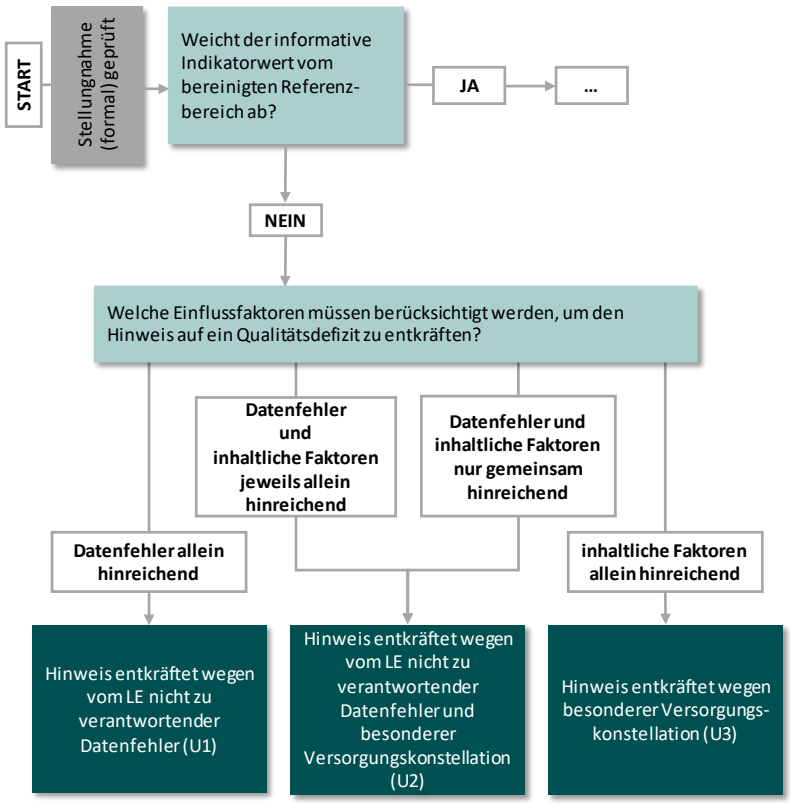

Abbildung 36: Algorithmus für die quantitative Differenzierung der Gründe für einen entkräfteten Hinweis auf ein Qualitätsdefizit

Die Differenzierung der Unterkategorien erfolgt in diesem Fall durch zwei weitere Varianten des informativen Indikatorwerts, bei denen der ursprüngliche Indikatorwert einmal nur um Datenfehler und einmal nur um fachlich-inhaltliche Faktoren bereinigt wurde. Die erforderlichen partiellen Nachberechnungen, die für die Prüfung der hier beschriebenen Bedingungen erforderlich sind, erfolgen analog zu dem in den vorangegangenen Abschnitten geschilderten Vorgehen.

Führt die partielle Nachberechnung bereits dann zu einem Indikatorergebnis im Referenzbereich, wenn ausschließlich Datenfehler, die nicht vom Leistungserbringer zu verantworten sind, bei der partiellen Nachberechnung berücksichtigt wurden, so ist die ursprüngliche Abweichung (des originalen Indikatorwerts vom Referenzbereich) hinreichend durch die Datenfehler erklärt. Führt zusätzlich die ausschließliche Berücksichtigung fachlich-inhaltlicher Einflussfaktoren, die nicht vom Leistungserbringer zu verantworten sind, bei der partiellen Nachberechnung ebenfalls zu einem Ergebnis innerhalb des Referenzbereichs, so ist die ursprüngliche Abweichung genauso hinreichend durch fachlich-inhaltliche Einflussfaktoren erklärt. Der ursprüngliche Hinweis wurde in diesem Fall also sowohl durch Datenfehler als auch durch eine besondere Versorgungskonstellation entkräftet (Kategorie U2). Andernfalls waren allein die Datenfehler für die Entkräftung des Hinweises entscheidend (Kategorie U1)[71].

[71] Dabei ist es unerheblich, ob zusätzliche Dokumentationsfehler vorgelegen haben, die der Leistungserbringer beeinflussen konnte.

Reicht die alleinige Berücksichtigung der nicht vom Leistungserbringer zu verantwortenden Datenfehler dagegen nicht aus, um ein Indikatorergebnis im Referenzbereich bei der partiellen Nachberechnung zu erhalten, so war für die Entkräftung des Hinweises eine Berücksichtigung fachlich-inhaltlicher Einflussfaktoren notwendig. Führt in diesem Fall bereits die alleinige Berücksichtigung fachlich-inhaltlicher Einflussfaktoren, die nicht vom Leistungserbringer zu verantworten sind, zu einem Ergebnis innerhalb des Referenzbereichs, so waren diese inhaltlichen Faktoren entscheidend (Kategorie U3). Andernfalls wurde der ursprüngliche Hinweis auf ein Qualitätsdefizit erst durch die kombinierte Berücksichtigung von Datenfehlern und besonderer Versorgungskonstellation entkräftet (Kategorie U2).

Differenzierung nach Entkräftung des Hinweises anhand der heuristischen Beurteilung

Wenn der indikatorbasierte Hinweis auf ein Qualitätsdefizit nicht durch die partielle Nachberechnung und den Vergleich mit dem bereinigten Referenzbereich, sondern erst im Rahmen der heuristischen Beurteilung entkräftet wurde, so kann auch der jeweilige Anteil von Datenfehlern und fachlich-inhaltlichen Faktoren am Einstufungsergebnis nicht zuverlässig quantifiziert werden. Die Einstufung in die passende Unterkategorie (U1, U2, U3) soll in diesem Fall auf Grundlage der Expertise von LAG bzw. Bundesstelle und beratenden Expertinnen und Experten (Fachkommission) vorgenommen werden (Abbildung 37).

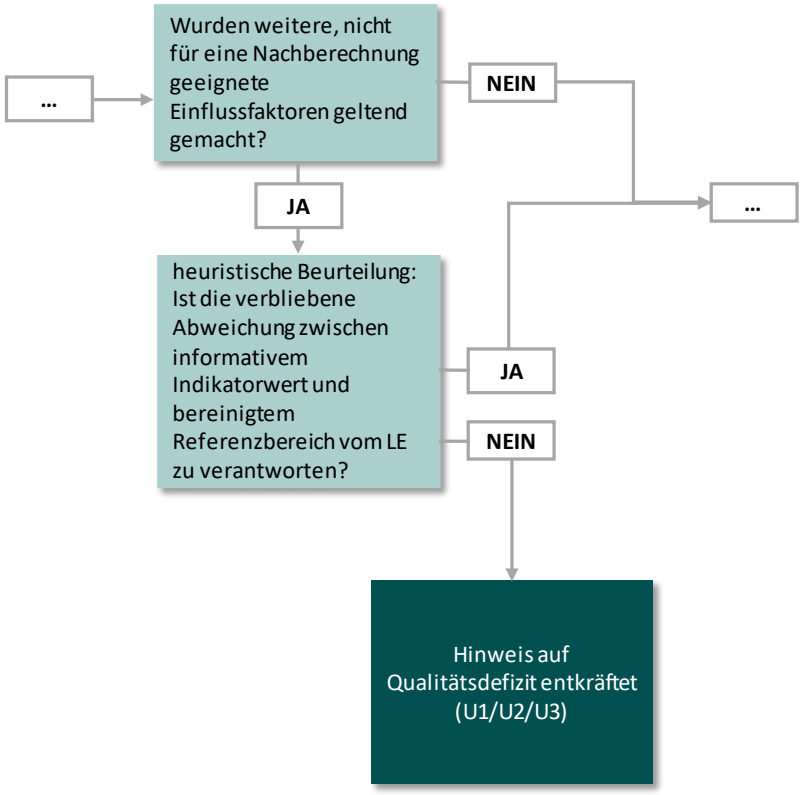

Abbildung 37: Heuristische Differenzierung der Gründe für einen entkräfteten Hinweis auf ein Qualitätsdefizits

6.5 Mindestanforderungen an Stellungnahmen

Weder in der QSKH-RL noch in der DeQS-RL finden sich konkrete Vorgaben für Form und Inhalt von Stellungnahmen sowie Angaben dazu, in welchem Zeitraum Stellungnahmen bei der LAG bzw. Bundesstelle abzugeben sind (vgl. Abschnitt 2.4). Dadurch ergibt sich ein großer Spielraum für die Ausgestaltung des Stellungnahmeverfahrens und damit eine mögliche Quelle für Heterogenität in den Ergebnissen, die nicht auf Qualitätsunterschiede zurückzuführen ist. Durch die Vorgabe von Kriterien für Stellungnahmen sollen diese formal und inhaltlich soweit wie möglich standardisiert werden und der Zeitraum für deren Erstellung beschränkt werden. Diese Standardisierung soll einerseits die Objektivität der Bewertung und andererseits eine Aufwandsreduktion bei der Bearbeitung der Stellungnahmen bewirken und bildet die Grundlage für einen einheitlichen Bewertungsprozess. Gerade das Einholen und die Bewertung von Präzisierungen zu Stellungnahmen (vgl. Abschnitt 2.4) stellen einen zusätzlichen Arbeitsaufwand dar. Mittels konkreter Kriterien für Stellungnahmen soll dieser Aufwand tendenziell verringert werden und eine weniger heterogene Bewertungsgrundlage (vgl. Abschnitt 3.3) erreicht werden. In die folgenden Empfehlungen sind u. a. die Rückmeldungen der Landesstellen im gemeinsamen Workshop mit dem IQTIG eingegangen (siehe Anhang, Kapitel 1).

6.5.1 Formale Kriterien

Die folgenden formalen Kriterien sollen bundeseinheitlich als Mindestanforderung für jede Stellungnahme gelten und von der LAG bzw. Bundesstelle geprüft werden. Es wird empfohlen, bundesweit einheitlich die Textvorschläge aus Anhang, Kapitel 7 für die Einholung von Stellungnahmen zu verwenden.

- Für jedes auffällige Indikatorergebnis ist eine Stellungnahme abzugeben.
- Die Stellungnahme ist innerhalb von 5 Wochen nach Anforderung der Stellungnahme abzugeben.
- Die Stellungnahme einschließlich eventuell notwendiger Anhänge ist in Schriftform zu erstellen und einzureichen.
- Belege, die ohne erläuternden Text übermittelt werden, sollen nicht als Stellungnahme anerkannt werden. Dazu zählen beispielsweise Einzelfalldokumente wie Arztbriefe, Epikrisen, Verlegungsberichte, Operationsberichte oder Befunde. Solche Belege können nur als Ergänzung einer Stellungnahme berücksichtigt werden und dazu in angemessenem Umfang vom Leistungserbringer eingereicht werden, um seine Argumentation zu unterstützen.
- Die fachlich verantwortliche Instanz des Leistungserbringers (z. B. die Chefärztin/der Chefarzt) ist über den Hinweis auf ein Qualitätsdefizit zu informieren und soll die angefertigte Stellungnahme autorisieren.[72]
- Die Anonymität sowohl der Patientinnen und Patienten (personenidentifizierende Daten) als auch ggf. weiterer Beteiligter (z. B. Zuweiser) ist zu wahren. Bei Nichteinhaltung der Vorgaben

[72] Anders als in der Richtlinie zu planungsrelevanten Indikatoren (plan. QI-RL), die festlegt, dass die juristisch verantwortliche Instanz die Stellungnahme zu autorisieren hat, wird die Autorisierung durch die juristisch verantwortliche Instanz für den Verwendungszweck der Qualitätsförderung oder für Public Reporting als zu aufwendig angesehen.

zur Wahrung der Anonymität soll die Stellungnahme aus Datenschutzgründen vernichtet werden und der Leistungserbringer über diesen Vorgang benachrichtigt werden.
- Der Leistungserbringer soll eine Erklärung darüber abgeben, ob die übermittelten QS-Daten, auf denen das Indikatorergebnis basiert, aus seiner Sicht korrekt sind.

Umgang mit der Nichteinhaltung formaler Vorgaben

Jedes Bewertungsverfahren benötigt Rahmenbedingungen, die die zulässigen Umstände der Bewertung vorgeben, sowie Regelungen für den Umgang mit der Nichteinhaltung solcher Vorgaben. Für das Modul „Qualitätsbewertung" sollen die obigen Kriterien Form und Inhalt von Stellungnahmen so weit wie möglich standardisieren und die empfohlenen Fristen für Datenlieferungen und für die Einreichung von Stellungnahmen etc. (siehe auch Abschnitt 8) sollen dem Modul einen zeitlichen Rahmen geben. Im Folgenden werden Empfehlungen für den Umgang mit der Nichteinhaltung dieser empfohlenen Vorgaben gegeben.

Gemäß dem Verständnis von Qualitätsindikatoren (vgl. Kapitel 3) gibt ein Indikatorergebnis außerhalb des Referenzbereichs einen Hinweis auf ein Qualitätsdefizit. Wird eine Stellungnahme nicht fristgerecht eingereicht oder erfüllt die eingereichte Stellungnahme die formalen Kriterien nicht (z. B. wegen Verstößen gegen den Datenschutz), kann sie nicht zur Entkräftung des Indikatorergebnisses herangezogen werden. Zwar ist eine Nicht-Einhaltung der formalen Vorgaben nicht gleichbedeutend mit einer Bestätigung der Indikatoraussage. Jedoch sollte es – im Sinne der Fairness – nicht möglich sein, durch nicht kriterienkonforme oder nicht fristgerechte Stellungnahmen die Qualitätsbewertung zu umgehen oder den Bewertungsprozess zu verzögern. Um Anreize für Formfehler zu vermeiden, wird daher empfohlen, in Fällen, in denen die formalen Vorgaben oder die Fristen nicht eingehalten wurden, die Bewertung „Qualitätsdefizit" zu vergeben. Dadurch wird deutlich, dass der betreffende Leistungserbringer keine adäquaten Informationen zur Verfügung gestellt hat, die die Indikatoraussage entkräften. In Fällen, in denen die Frist für die Einreichung einer Stellungnahme noch nicht verstrichen ist, soll dem Leistungserbringer die Möglichkeit gegeben werden, eine den Kriterien entsprechende Stellungnahme einzureichen.

6.5.2 Inhaltlicher Fokus

In den Stellungnahmen können für das Verfehlen des Referenzbereichs Gründe angegeben werden, die nur auf einzelne oder wenige Behandlungsfälle zutreffen (z. B. dass ein Notfall ein Abweichen vom üblichen Vorgehen erforderlich machte) oder die sich auf alle Behandlungsfälle beziehen (z. B. dass wegen einer Spezialisierung der Einrichtung auf bestimmte Behandlungen die Patientinnen und Patienten ein besonderes Risikoprofil haben). Vor dem Hintergrund der folgenden Argumente sollte die Stellungnahme in beiden Fällen möglichst auf eine Behandlungsfall-übergreifende Analyse fokussieren:

- Qualitätsindikatoren treffen eine aggregierte Qualitätssaussage auf Ebene des Leistungserbringers und dienen nicht dazu, Einzelfälle zu bewerten. Eine zu stark an einzelnen Fällen orientierte Stellungnahme kann den Blick auf systemische Ursachen und notwendige

systemische Verbesserungsansätze, wie z. B. die Verbesserung in der Einrichtungsorganisation oder der Patientensicherheitskultur, versperren.

- Das Ergebnis der fachlichen Bewertung soll keine umfassende Qualitätsaussage treffen (siehe Abschnitt 3.3), sondern nur die Angemessenheit der Indikatoraussage untersuchen. Ob eine besondere, nicht vom Leistungserbringer zu verantwortende Konstellation vorgelegen hat, lässt sich ggf. anhand weniger, auch zusammenfassender Informationen beurteilen, ohne alle betreffenden Behandlungsfälle im Detail zu prüfen. Beispielsweise ist es denkbar, dass ein Leistungserbringer, dessen In-Hospital-Sterblichkeit von Patientinnen und Patienten mit Pneumonie außerhalb des Referenzbereichs liegt, in seiner Stellungnahme die räumliche Nähe zu einer Hospiz-Einrichtung als unberücksichtigten Risikofaktor anführen möchte. Als Beleg könnte in diesem Fall z. B. eine Übersicht dienen, welche der Patientinnen und Patienten im Nenner des Indikators mit palliativer Behandlungsintention aufgenommen wurden, ggf. ergänzend belegt durch die jeweiligen Nebendiagnosen. Eine ausführliche Untersuchung aller Behandlungsverläufe anhand von Epikrisen wird für die Beurteilung, ob der Indikator das Qualitätsmerkmal (Sterblichkeit an Pneumonie bei kurativer Behandlungsintention) angemessen für diesen Leistungserbringer abbildet, voraussichtlich nicht benötigt.
- Umfangreiche Diskussionen von Einzelfällen in den Stellungnahmen gehen mit einem hohen Aufwand bei Stellungnehmenden und LAG bzw. Bundesstelle einher. Dazu trägt bei, dass schriftliche Epikrisen sehr detailliert sein müssen und potenzielle Fragen der beurteilenden Fachkommission bereits vorwegnehmen müssen, wenn der Leistungserbringer die Nachvollziehbarkeit seiner Argumentation sicherstellen möchte. Die Erkenntnisse, die sich aus detaillierten Einzelfallbetrachtungen ergeben, sind jedoch weniger für die Bewertung der Indikatoraussage wichtig (s. o.), sondern vielmehr für die Identifizierung von Ansätzen und Maßnahmen zur Qualitätsverbesserung hilfreich. Detaillierte Einzelfallbetrachtungen sollten daher möglichst dem Modul „Qualitätsförderung“ vorbehalten bleiben. Dies führt auch zu einer Aufwandsreduktion, erstens, weil in den Fällen, in denen keine detaillierten Informationen für die Qualitätsbewertung benötigt werden und in denen letztendlich keine Qualitätsförderung notwendig ist, auf die Anfertigung ausführlicher Epikrisen für die externe QS und deren Diskussion verzichtet wird. Zweitens können im Rahmen der Qualitätsförderung die relevanten Fragestellungen durch den Einsatz echter dialogischer Elemente (z. B. Gespräche und Visitationen) zielgerichteter beantwortet werden als durch ein schriftliches Verfahren.

Bei kleinen Fallzahlen oder kleinen Zahlen von interessierenden Ereignissen, die zu einem Hinweis auf ein Qualitätsdefizit geführt haben, ist es einem Leistungserbringer oft nicht möglich, die dem Problem zugrunde liegenden Systemfaktoren zu identifizieren und eine Behandlungsfall-übergreifende Analyse vorzunehmen. Die unter Abschnitt 6.4 geschilderte Methodik erlaubt es daher, sowohl fallbezogene als auch fallübergreifende Gründe für das Verfehlen des Referenzbereichs bei der fachlichen Bewertung zu berücksichtigen.

Für die Prüfung der Indikatoraussage in der fachlichen Bewertung ist es wichtig, dass der Leistungserbringer in seiner Stellungnahme für jeden Grund angibt, auf welche seiner Behandlungsfälle dieser Grund zutrifft bzw. ob der Grund für alle Behandlungsfälle zutrifft. Dazu

sind die Vorgangsnummern der entsprechenden Behandlungsfälle mit der Stellungnahme zu übermitteln. Diese Angabe ist auch für diejenigen Behandlungsfälle zu machen, bei denen kein interessierendes Ereignis aufgetreten ist. Für dieses Vorgehen gibt es zwei Gründe:

- Zum einen sollte ein Behandlungsfall, für den das Qualitätsziel des Indikators nicht anwendbar ist, da ein nicht vom Leistungserbringer zu verantwortender Einflussfaktor vorliegt, nicht zur Qualitätsbeurteilung herangezogen werden, unabhängig davon, ob dieser Behandlungsfall ein interessierendes Ereignis aufgewiesen hat oder nicht. Beispielsweise könnte ein Leistungserbringer eine Stellungnahme einreichen zu einem Qualitätsindikator zur Durchführung einer postoperativen Thromboseprophylaxe, der den Anteil der Patientinnen und Patienten mit erhaltener Prophylaxe misst und einen Referenzbereich von > 90 % hat. Wenn der Leistungserbringer für die 50 von ihm im Erfassungsjahr behandelten Patientinnen und Patienten einen Indikatorwert von 80 % aufweist (d. h. in 40 der 50 Fälle wurde eine Prophylaxe durchgeführt) und er geltend machen möchte, dass bei einem besonders hohen Anteil der Patientinnen und Patienten eine Kontraindikation zur blutverdünnenden Behandlung vorlag, reicht es nicht aus, das Vorliegen einer Kontraindikation in z. B. 6 der 10 Fälle zu belegen, die bei der Indikatorberechnung als interessierendes Ereignis (keine Thromboseprophylaxe) gewertet wurden. Es bliebe in diesem Fall unbekannt, ob alle verbleibenden 44 Fälle des Leistungserbringers für eine Thromboseprophylaxe in Frage kamen oder weniger als 44. Im ersten Fall betrüge der nachberechnete Indikatorwert 40 von 44 Fällen (90,9 %). Im anderen Fall wäre denkbar, dass die anerkannte Kontraindikation z. B. auch bei 15 der 40 Fälle ohne interessierendes Ereignis (d. h. mit durchgeführter Thromboseprophylaxe) vorlag. Die Zielpopulation des Indikators (Fälle, bei denen eine Thromboseprophylaxe durchgeführt werden soll) wäre dann nur 29 Fälle groß (50 Fälle abzgl. 21 Fällen mit Kontraindikation), bei denen der Leistungserbringer in 4 Fällen nicht begründen konnte, warum keine Thromboseprophylaxe durchgeführt wurde (nachberechneter Indikatorwert 86,2 %). Liegen Informationen über das Vorliegen eines Einflussfaktors nicht für alle Behandlungsfälle vor, kann dies also zu relevanten Unterschieden führen.
- Zum anderen kann es vorkommen, dass mehrere Einflussfaktoren auf das Indikatorergebnis von einem Leistungserbringer geltend gemacht und bei der fachlichen Bewertung als nicht von ihm zu verantworten anerkannt werden. In diesen Fall muss anhand der Vorgangsnummern erkennbar sein, in welcher Kombination die anerkannten Einflussfaktoren bei den Behandlungsfällen vorliegen, damit die Zielpopulation bei der Nachberechnung des Indikatorwert korrekt bestimmt werden kann.

In der Berechnung des informativen Indikatorergebnisses sind diese Überlegungen berücksichtigt (siehe Abschnitte 6.4.1 und Anhang, Kapitel 6). In Kapitel 7 des Anhangs findet sich ein Beispiel, auf welche Weise in einer Stellungnahme die Vorgangsnummern fallbezogen übermittelt werden sollten. Es wird empfohlen, eine digitale Übermittlung der Vorgangsnummern festzulegen, um Übertragungsfehler in die Software für die Nachberechnung des informativen Indikatorwerts (siehe Abschnitt 6.6) zu vermeiden. Auf die Angabe der Vorgangsnummern soll nur dann verzichtet werden, wenn dies nicht sinnvoll oder nicht möglich ist, beispielsweise weil ein struktureller Einflussfaktor (etwa die Anfahrtszeit des Rettungsdienstes in einer ländlichen Region, siehe Beispiel 3 in Kapitel 5 des Anhangs) sich auf

alle Behandlungsfälle gleichermaßen bezieht. Diese Einflussfaktoren können bei der fachlichen Bewertung im heuristischen Beurteilungsschritt (siehe Abschnitt 6.4) berücksichtigt werden. Ein Einflussfaktor, bei dem aus der Stellungnahme nicht hervorgeht, in welchem Maße er auch bei Fällen ohne interessierendes Ereignis vorgelegen hat, soll in der fachlichen Bewertung nicht berücksichtigt werden.

Werden in der Stellungnahme Gründe für das auffällige Indikatorergebnis angeführt, die aus Sicht des Leistungserbringers nicht von ihm zu verantworten waren, so sollen diese Gründe nicht nur benannt werden, sondern es soll auch nachgewiesen werden, dass mit der gebotenen Sorgfaltspflicht alles Zumutbare getan wurde, um diese aufzufangen (z. B. die Einplanung von genügend Ressourcen und genügend Zeit). Damit ist nicht gemeint, dass jeder Behandlungsfall in der Stellungnahme ausführlich diskutiert wird (s. o.), sondern dass z. B. anhand einer fallübergreifenden Erläuterung für die Fachkommission nachvollziehbar werden muss, warum der geltend gemachte Einflussfaktor tatsächlich nicht vom Leistungserbringer zu verantworten war. Das gleiche trifft sinngemäß auf Datenfehler zu: Hat ein Leistungserbringer in seiner Stellungnahme angegeben, dass Fehler in den übermittelten QS-Daten vorliegen, soll die Stellungnahme erklären, welche Fehler seiner Ansicht nach vorliegen und welche der Fehler in den übermittelten Daten von ihm nicht zu vertreten sind.

Analog zu den formalen Vorgaben für die Stellungnahmen (siehe Abschnitt 6.5.1) liegt auch die Verantwortung für die Vollständigkeit und Angemessenheit der Inhalte einer Stellungnahme beim Leistungserbringer. Aus Gründen der Fairness und der Verfahrenseffizienz sollte auf die Nachforderung von ergänzenden Informationen durch die LAG verzichtet werden (siehe Abschnitt 8.2.2). Kann innerhalb der Einreichungsfrist ein Sachverhalt, der mutmaßlich die Indikatoraussage entkräftet, nicht hinreichend durch entsprechende Informationen und Belege plausibel gemacht werden, so stellt der indikatorbasierte Hinweis auf ein Qualitätsdefizit die bestverfügbare Evidenz bzgl. der Erfüllung des Qualitätsmerkmals dar. Die fachliche Bewertung soll dann auf Basis der bis zu diesem Zeitpunkt vorliegenden Informationen erfolgen.

6.6 Umsetzung und Aufwand

Das in den vorangehenden Abschnitten empfohlene Konzept für die fachliche Bewertung zeichnet sich gegenüber dem bisherigen Vorgehen durch mehr explizite Verfahrensregeln und einem höheren Grad an Standardisierung aus. Dies kann zu Herausforderungen bei der Implementation und zu höherem Aufwand bei der Durchführung als bisher führen. Diese Aspekte werden im Folgenden diskutiert.

Klassifikation der Einflussfaktoren

Die initialen Schritte bei der fachlichen Bewertung bestehen in der Klassifikation der in den Stellungnahmen angeführten Einflussfaktoren auf das Indikatorergebnis (siehe Abschnitte 6.3 und 6.4.1). Sie ist eine Voraussetzung für die Beurteilung, ob nicht berücksichtigte Einflussfaktoren die Indikatoraussage entkräften. Wenn auch im bisherigen Strukturierten Dialog der Einfluss solcher Faktoren auf das Indikatorergebnis in die Beurteilung eingeht, führt das vorgeschlagene Vorgehen zu keinem höheren Aufwand, da in diesem Fall auch bisher schon eine solche

Klassifikation vorgenommen werden musste. Wenn im bisherigen Strukturierten Dialog ein anderes Ziel verfolgt wird (z. B. eine Gesamtschau der Behandlungsqualität eines Leistungserbringers), kann sich der Aufwand bei den LAG und den Fachkommissionen gegenüber dem Status quo unterscheiden. Die Erfahrungen mit dem Vorgehen im Verfahren zu planungsrelevanten Qualitätsindikatoren zeigen, dass eine explizite Klassifikation der in Stellungnahmen angeführten Einflussfaktoren auf das Indikatorergebnis grundsätzlich umsetzbar ist.

Partielle Nachberechnung

Die Nachberechnung des informativen Indikatorwerts auf Grundlage der anerkannten Einflussfaktoren ist ein wesentliches Element des vorgeschlagenen Algorithmus für die fachliche Bewertung und im Vergleich zum bisherigen Vorgehen neu. Die grundsätzliche Umsetzbarkeit dieses Vorgehens zeigen die Erfahrungen aus dem Verfahren der planungsrelevanten Qualitätsindikatoren. Eine zuverlässige und aufwandsangemessene Umsetzung dieses Elements setzt geeignete technische Unterstützung voraus, beispielsweise in Form eines für alle Bewertungsstellen einheitlichen Web-basierten Portals. Eine einheitliche Softwarelösung für alle Bewertungsstellen würde zudem Heterogenität in der Bewertung aufgrund unterschiedlicher Softwarelösungen bei den Landesstellen vermeiden. Die Ergebnisse der Nachberechnungen sollten von der Software im Echtzeit-Betrieb bereitgestellt werden, sodass während der Beratungen in den Fachkommissionssitzungen keine Verzögerungen entstehen. Zu berücksichtigen ist der Aufwand für die Implementation und Pflege einer solchen Software, der bei einer zentralen technischen Lösung z. B. durch die Schaffung des erforderlichen Rechtesystems und von Berechnungsmöglichkeiten in Echtzeit entsteht. Diese notwendige IT-seitige Entwicklungsarbeit muss bei einer möglichen Beauftragung der Umsetzung des Konzeptes berücksichtigt werden.

Nach der Klassifikation der in einer Stellungnahme angeführten Einflussfaktoren (s. o.) besteht der zusätzliche Aufwand durch die hier empfohlene Methodik darin, der Software die Behandlungsfälle z. B. in Form einer Liste (siehe Kapitel 8 des Anhangs) anzugeben, die bei der Nachberechnung aus der Indikatorberechnung ausgeschlossen werden sollen. Dazu müssen diese Angaben vom Leistungserbringer zur Verfügung gestellt werden. Möchte ein Leistungserbringer einen bisher nicht erfassten Einflussfaktor geltend machen, muss er – sofern nicht alle Behandlungsfälle betroffen sind – die Vorgangsnummern aller Behandlungsfälle mit diesem Einflussfaktor übermitteln (siehe Abschnitt 6.5.2). Im Vergleich zum bisher üblichen Vorgehen, bei dem die Angabe üblicherweise nur für Behandlungsfälle mit interessierendem Ereignis erfolgt, ergibt sich für die Leistungserbringer ein Mehraufwand. Dieser Mehraufwand ist nicht vermeidbar, wenn eine Verzerrung durch die selektive Betrachtung von bestimmten Fällen vermieden werden soll.

Differenzierung mittels Ziffern für Qualitätsdefizit

Zur Erhöhung der Transparenz über die Bewertungsergebnisse für die Verfahrensbeteiligten wurden im Bewertungsschema auf der zweiten Ebene Ziffern und ein aus zwei Schritten bestehender Bewertungsalgorithmus vorgeschlagen (siehe Abschnitte 6.1.3 und 6.4.4). Eine Software für die Nachberechnung des informativen Indikatorwertes kann den ersten, expliziten Beurtei-

lungsschritt für diese Differenzierung ohne zusätzlichen Aufwand für die LAG bzw. Fachkommission vornehmen, da er auf denselben Informationen basiert. Der zweite, heuristische Beurteilungsschritt bedeutet im Vergleich zum bisherigen Vorgehen ebenfalls keinen Zusatzaufwand, da eine Beurteilung von Dokumentationsproblemen auch im bisherigen Verfahren schon durchgeführt wird, wie sich an der Vergabe entsprechender Schlüsselwerte zeigt (z. B. in der Kategorie „D“ des bisherigen Bewertungsschemas, siehe Abschnitt 2.4.3).

Differenzierung mittels Ziffern für „kein Hinweis auf Qualitätsdefizit“

Für die Unterscheidung der Gründe, weshalb der Hinweis aus einem Indikatorergebnis als entkräftet angesehen wurde, wurden ebenfalls Ziffern und Bewertungsregeln vorgeschlagen (siehe Abschnitte 6.1.2 und 6.4.5). Erfolgte die Entkräftung auf Grundlage der Nachberechnung des informativen Indikatorwerts, so kann die passende Ziffer ohne zusätzlichen Aufwand für die LAG bzw. Fachkommission von einer entsprechenden Software mit ausgegeben werden, da alle notwendigen Informationen bereits angegeben wurden. Wurde der Hinweis auf ein Qualitätsdefizit durch heuristische Beurteilung entkräftet, so waren entweder nicht zu verantwortende Datenfehler, eine besondere Versorgungskonstellation oder eine Kombination aus beidem ausschlaggebend für die Entkräftung. Die Differenzierung der Ziffern ist dann lediglich eine Dokumentation der Entscheidungsgründe für die getroffene Beurteilung und bedeutet keinen Mehraufwand.

Trennung von Qualitätsbewertung und -förderung

Die Prozesse von Qualitätsförderung und Qualitätsbewertung sollen zukünftig getrennt werden, auch um die Objektivität der Bewertung zu erhöhen. Für die Qualitätsförderung können jedoch über das bloße Einstufungsergebnis hinaus weitere Informationen aus der fachlichen Bewertung relevant sein. Beispielsweise können sich aus Fallberichten, die vom Leistungserbringer als Beleg für eine besondere Versorgungskonstellation angeführt wurden, Hinweise ergeben, durch welche Maßnahmen der Leistungserbringer Qualitätsverbesserungen erreichen kann. Damit diese Informationen bei der Qualitätsförderung genutzt werden können, müssen diese bei der fachlichen Bewertung neben der Qualitätsbewertung im eigentlichen Sinne begleitend dokumentiert werden. Wenn die Qualitätsförderung zu einem späteren Zeitpunkt als die Qualitätsbewertung erfolgt (z. B. nicht in derselben Sitzung der Fachkommission), so bedeutet bereits die notwendige erneute Befassung einen gewissen Aufwand. Dieser entsteht allerdings ebenfalls im bisherigen Vorgehen, wenn Qualitätsverbesserungsmaßnahmen wie kollegiale Gespräche und Visitationen zu einem späteren Zeitpunkt als die initiale Beurteilung der schriftlichen Stellungnahme durchgeführt werden. Dies dürfte auch für die in § 17 Abs. 3 DeQS-RL genannten Maßnahmen der Qualitätsförderung zutreffen. Maßnahmen, die allein auf Basis der Stellungnahme erfolgen (beispielsweise die Zusendung eines indikatorspezifischen Fragebogens zur Selbsteinschätzung des Leistungserbringers), können dagegen noch innerhalb derselben Sitzung der Fachkommission veranlasst werden. Dies ist auch im vorgeschlagenen Konzept möglich.

Zusammenfassung

Durch den vorgeschlagenen, auf möglichst expliziten Beurteilungsregeln basierenden Bewertungsprozess entstehen voraussichtlich folgende zusätzliche Aufwände gegenüber einem ausschließlich heuristischen, auf impliziten Beurteilungsregeln basierenden Vorgehen:

- Bereitstellung und Pflege einer technischen Unterstützung (Software) von Nachberechnungen
- Eingabe der relevanten Information in die Software bei der fachlichen Bewertung
- Nochmalige Befassung mit Informationen aus den Stellungnahmen bei von der Qualitätsbewertung zeitlich getrennten Befassung der Fachkommission mit Maßnahmen der Qualitätsverbesserung.

Eine Reihe weiterer Aufwände bei der Qualitätsbewertung sind dagegen nicht spezifisch für das vorgeschlagene Vorgehen, sondern sind in jedem Fall zu berücksichtigen, wenn fundierte Aussagen über die Erfüllung des Qualitätsmerkmals eines Indikators gemäß dem in Abschnitt 6.1 dargestellten Bewertungsschema getroffen werden sollen:

- Klassifikation der in den Stellungnahmen genannten Einflussfaktoren hinsichtlich Verantwortlichkeit, Art des Faktors und Auswirkung auf das Indikatorergebnis
- Angabe durch den Leistungserbringer, welche Behandlungsfälle von einem geltend gemachten Einflussfaktor betroffen sind
- Beurteilung, inwieweit die Abweichung eines Indikatorergebnisses vom Referenzbereich auf Datenfehler zurückzuführen ist, sofern sich dies nicht berechnen lässt
- Beurteilung, ob sich bereits auf Grundlage der schriftlichen Stellungnahme Hinweise auf Maßnahmen zur Qualitätsverbesserung ableiten lassen

Eine Reduktion des Aufwandes für die fachliche Bewertung könnte in geringem Maße durch Verzicht auf die Differenzierung der Ziffern des Bewertungsschemas erreicht werden. Allerdings ist die Vergabe dieser Ziffern sowohl bei einem expliziten als auch bei einem heuristischen Vorgehen im Wesentlichen ein Nebenprodukt des Beurteilungsprozesses und erzeugt wenig Zusatzaufwand (s. o.).

6.7 Zusammensetzung von Fachkommissionen

Die Fachkommissionen haben eine wichtige beratende Funktion für die LAG bzw. Bundesstelle bei der Qualitätsbewertung und Qualitätsförderung. Ihre Arbeit verfolgt im Rahmen des Stellungnahmeverfahrens zwei Ziele. Zum einen soll die Fachkommission die Angemessenheit der Abbildung der Versorgungsqualität anhand des Indikators aus medizinisch-fachlicher Sicht beurteilen. Dafür müssen die Mitglieder der Fachkommission die indikatorbezogenen Stellungnahmen lesen und eine fachliche Beurteilung vornehmen. Die fachliche Beurteilung beinhaltet die Prüfung, ob eine besondere Versorgungskonstellation vorliegt, die zur Fehlmessung im Indikator geführt hat, sowie eine Empfehlung an die LAG bzw. Bundesstelle, wie das Ergebnis in der abschließenden Qualitätsbewertung kategorisiert werden sollte (siehe Abschnitte 6.3 und 6.4). Zum anderen soll die Fachkommission nach dem Bewertungsprozess und bei einem bestätigten

Qualitätsdefizit Empfehlungen darüber abgeben, welche Maßnahmen der LAG bzw. Bundesstelle zur Qualitätsverbesserung an das interne Qualitätsmanagement adressiert werden.

Orientiert an den Zielen und Aufgaben der Fachkommission sind bestimmte Anforderungen sowohl an die einzelnen Expertinnen und Experten als auch an die Zusammensetzung der Gruppe als Ganzes zu formulieren. Diese sollten insbesondere die in Abschnitt 3.7 beschriebenen Erfolgsfaktoren (fachliche Kompetenz, Unabhängigkeit und Interdisziplinarität) bei der Zusammensetzung der Fachkommission berücksichtigen. Daher wird empfohlen, eine fachlich qualifizierte, unabhängige und interdisziplinäre Fachkommission zusammenzusetzen, um die Akzeptanz des Gesamtverfahrens sicherzustellen und Qualitätsdefizite erfolgversprechend abzubauen. Die Verantwortung für die Zusammensetzung einer fachlich qualifizierten, unabhängigen und interdisziplinären Fachkommission soll bei der LAG bzw. Bundesstelle liegen, wobei die im Folgenden dargestellten, bundeseinheitlichen Kriterien für die Auswahl einzelner Expertinnen und Experten und für die Zusammensetzung der Fachkommission als Rahmenbedingungen vorgegeben werden sollten. Zur Steigerung der Akzeptanz des Verfahrens gegenüber den Expertinnen und Experten der Fachkommission wird eine bundesweit einheitliche Regelung zur Aufwandsentschädigung, die gleichermaßen für alle Expertinnen und Experten gilt, empfohlen. Auch im IQTIG-LQS-Treffen am 18. Juni 2019 wurde eine einheitliche Aufwandsentschädigung seitens einiger Teilnehmerinnen und Teilnehmer befürwortet (siehe Anhang, Kapitel 2). In Abschnitt 6.7.1 werden zunächst die Kriterien beschrieben, die von den einzelnen Expertinnen und Experten persönlich erfüllt sein sollten, während in Abschnitt 6.7.2 erläutert wird, wie die Fachkommission als Gruppe zusammengesetzt sein sollte.

6.7.1 Kriterien für die Auswahl der Expertinnen und Experten

Bei den Kriterien für die Auswahl der einzelnen Expertinnen und Experten kann zwischen fachlich-inhaltlichen Anforderungskriterien, die sich auf die fachliche Eignung (Qualifikation und berufliche Erfahrung) beziehen, und formalen Anforderungskriterien (z. B. Zusicherung der Vertraulichkeit, Angaben zu Interessenkonflikten) unterschieden werden. Die folgenden Anforderungskriterien sollen sowohl die fachliche Kompetenz als auch den Einbezug möglichst unabhängiger Mitglieder der Fachkommission gewährleisten. Es wird empfohlen, die Kriterien erst ab der nächsten Neubesetzung der Fachkommissionen anzulegen.

6.7.1.1 Fachlich-inhaltliche Anforderungskriterien

Um eine möglichst nachvollziehbare Beurteilung der besonderen Versorgungskonstellationen vornehmen zu können, sollten die Fachexpertinnen und -experten nachweislich über die für das jeweilige QS-Verfahren relevante klinisch-praktische Erfahrung verfügen. Die fachliche Expertise und die Kenntnis der Versorgungspraxis ist dabei eine notwendige Voraussetzung für die Beurteilung der in den Stellungnahmen vorgebrachten Gründe und die Bewertung dahingehend, ob es sich dabei um eine besondere Versorgungskonstellation handelt. Angesichts des sich schnell entwickelnden medizinischen Wissens und der Vielzahl an existierenden, fachlich unterschiedlichen Versorgungsbereichen ist die LAG bzw. Bundesstelle darauf angewiesen, sich von Expertinnen und Experten beraten zu lassen, deren Wissen und Erfahrungen möglichst aktuell sind. Die

LAG bzw. Bundesstelle sollte die jeweiligen Fachgesellschaften über eine ausstehende Neubenennung der Mitglieder einer Fachkommission informieren, damit eine Bewerbung von Mitgliedern der entsprechenden Fachgesellschaften möglich ist. Die zu berufenden ärztlichen Fachexpertinnen und -experten sollten nach der Facharztausbildung während der letzten drei Jahre vor ihrem Mitwirken in der Fachkommission im entsprechenden Versorgungsbereich praktisch tätig gewesen sein. Auch nicht medizinische und an der Leistungserbringung beteiligte Fachexpertinnen und -experten (z. B. Pflegekräfte, Logopädinnen/Logopäden, Ergotherapeutinnen/Ergotherapeuten, Physiotherapeutinnen/Physiotherapeuten) sollten nach ihrer qualifizierenden Ausbildung während der letzten drei Jahre im entsprechenden Versorgungsbereich tätig gewesen sein. Expertinnen und Experten, die während ihrer Mitgliedschaft in der Fachkommission in den Ruhestand wechseln oder eine Tätigkeit außerhalb der Gesundheitsversorgung ausüben, können bis zu vier Jahre nach ihrem Ausscheiden aus der Tätigkeit in der Gesundheitsversorgung mitwirken. Expertinnen und Experten, die im Ausland tätig sind oder waren, sollten mit den nationalen fachlichen Standards vertraut sein und mindestens fünf Jahre als Klinikerinnen/Kliniker in Deutschland tätig gewesen sein. Bei längeren Auslandsaufenthalten ist die Eignung der Expertinnen und Experten im Einzelfall abzuwägen. Um Empfehlungen für qualitätsverbessernde Maßnahmen abgeben zu können, sind Kenntnisse der Qualitätssicherung und des internen Qualitätsmanagements hilfreich.

6.7.1.2 Formale Anforderungskriterien

Eine Bewerbung sollte unter Übermittlung folgender Unterlagen (Bewerbungsunterlagen) an die LAG bzw. Bundesstelle erfolgen:

- Aussagekräftiger Lebenslauf (mit notwendigen Fachkenntnisse und -erfahrungen, ggf. auch in den Bereichen Qualitätssicherung und Qualitätsmanagement)
- Offenlegung von Interessenkonflikten (inkl. Angaben zu Gremienarbeit)
- Einwilligung, dass Interessenkonflikte ggf. bei Berufung öffentlich genannt werden
- Verpflichtungserklärung zur Wahrung der Vertraulichkeit
- Zusicherung einer regelmäßigen Beteiligung an den Sitzungen der Fachkommission
- Anerkennung der Geschäftsordnung des jeweiligen Gremiums
- Bei Angestelltenverhältnis: Zustimmungserklärung des Arbeitgebers

Lebenslauf

Anhand des Lebenslaufs sind die notwendigen Fachkenntnisse und Erfahrungen zu prüfen. Nur bei Erfüllung der beschriebenen fachlich-inhaltlichen Kriterien darf eine Expertin oder ein Experte ausgewählt werden. Werden Expertinnen und Experten trotz unzureichender Erfüllung der fachlich-inhaltlicher Anforderungen von der LAG bzw. Bundesstelle benannt, sind die Gründe für die Benennung schriftlich niederzulegen.

Interessenkonflikte

Bei Akteuren im Gesundheitswesen können primäre und sekundäre Interessen das Handeln beeinflussen und im Konflikt zueinanderstehen. Das primäre Interesse im Rahmen des Stellung-

nahmeverfahrens sollte die sachgemäße Bewertung der Versorgungsqualität anhand der Indikatoren sein. Damit eine unabhängige Bewertung der Versorgungsqualität gewährleistet werden kann, müssen Konflikte mit sekundären Interessenslagen weitestgehend vermieden werden. Solche Interessenkonflikte können beispielsweise auffällige Indikatorergebnisse der eigenen Einrichtung sein oder eine Konkurrenzsituation geografisch benachbarter Einrichtungen.

Ein wesentlicher Aspekt des Umgangs mit Interessenkonflikten ist deren Transparenz. Daher müssen Fachexpertinnen und -experten bereit sein, dass ihre Interessenkonflikte im Vorfeld einer Berufung zumindest dem Lenkungsgremium des Verfahrens in der LAG bzw. Bundesstelle offengelegt werden. Können Interessenkonflikte nicht ausreichend vermieden werden, sind sie ggf. zu veröffentlichen. Die Prüfung, Einordnung und Veröffentlichung der Interessenkonflikte liegt in der Verantwortung der LAG bzw. Bundesstelle.

Vertraulichkeit

Die sich bewerbenden Expertinnen und Experten müssen sich zur strikten Wahrung der Vertraulichkeit von Unterlagen und Diskussionsinhalten verpflichten. Ein Verstoß dagegen soll zum Ausschluss mit sofortiger Wirkung führen. Eine Fachkommission kann bei Missbrauch der Vertraulichkeit von der LAG bzw. Bundesstelle aufgelöst werden.

Zeitliche Verfügbarkeit

Die sich bewerbenden Expertinnen und Experten müssen im Vorfeld schriftlich bestätigen, dass sie zeitlich in der Lage sind, die Teilnahme an den Sitzungen sowie deren Vor- und Nachbereitung zu gewährleisten. Bei abhängig beschäftigten Personen ist eine Bestätigung des Arbeitgebers beizufügen, in der erklärt wird, dass der Arbeitgeber die Wahrnehmung dieses Amtes unterstützt und der Mitarbeiterin / dem Mitarbeiter die Teilnahme an den Sitzungen der Fachkommission ermöglichen wird. Dieses Kriterium ist zur Aufrechterhaltung einer konstanten Sitzungsbeteiligung insbesondere unter Einbeziehung von beruflich tätigen Mitgliedern notwendig.

Weitere Gründe, Mitteilung bei Änderung der Gegebenheiten

Weiterhin hat eine bewerbende Person zu versichern, dass zusätzlich keine anderen Gründe vorliegen, die der Fachlichkeit oder dem Ansehen des Verfahrens entgegenstehen könnten. Ändert sich nach Berufung die Situation eines Mitglieds der Fachkommission so, dass ein neuer oder verstärkter Interessenkonflikt angenommen werden kann oder dass ein anderer relevanter Grund eintritt, so muss diese Änderung der LAG bzw. Bundesstelle gegenüber unverzüglich, spätestens aber zwei Wochen vor der nächsten Beratungsleistung, angezeigt werden. Die Nicht-Anzeige eines solchen Umstands kann von der LAG bzw. Bundesstelle als alleiniger Grund zum Anlass genommen werden, um eine Expertin bzw. einen Experten von der weiteren Mitwirkung fristlos auszuschließen.

6.7.2 Kriterien für die Zusammensetzung der Fachkommission

Im Vordergrund für die Zusammensetzung der Fachkommission steht der Erfolgsfaktor der *Interdisziplinarität*. Interdisziplinarität bedeutet im Kontext des Stellungnahmeverfahrens, dass

für eine sachgerechte Bewertung der Versorgungsqualität anhand der Indikatoren und für Empfehlungen zur Qualitätsförderung alle an den Versorgungergebnissen beteiligten Berufsgruppen einbezogen werden sollen, die auf die Indikatorergebnisse wesentlichen Einfluss haben. Um die Interdisziplinarität zu wahren, sollte vor dem Hintergrund der einzelnen, sehr unterschiedlichen Versorgungsbereiche daher die Zusammensetzung der jeweiligen Fachkommission spezifisch ausgewählt werden. Dabei ist für jeden Versorgungsbereich zu identifizieren, welche Berufsgruppen durch ihre Versorgungsleistung einen wesentlichen Einfluss auf die Ergebnisse der Qualitätsindikatoren ausüben. Beispielsweise hat im QS-Verfahren *Geburtshilfe* die Berufsgruppe der Hebammen und Geburtshelfer einen Einfluss auf die Ergebnisse der Indikatoren zum Dammriss. Durch die Einbindung relevanter Berufsgruppen soll einerseits vermieden werden, dass die Bewertung der Versorgungsqualität durch die Interessen einzelner Berufsgruppen verzerrt wird. Andererseits können durch eine interdisziplinäre Fachkommission berufsgruppenübergreifende Probleme identifiziert und entsprechende Fördermaßnahmen empfohlen werden. In Abschnitt 3.7 wurde zudem dargelegt, dass dabei auch Patientinnen und Patienten einbezogen werden müssen.

Für eine interdisziplinäre Fachkommission wird empfohlen, folgende **ständige Expertinnen und Experten** für jeweils alle QS-Verfahren zu benennen:

Fachärztinnen und Fachärzte

Fachärztinnen und Fachärzte tragen in der Regel die Hauptverantwortung für das Behandlungsergebnis. Daher sind mindestens zwei Fachärztinnen bzw. Fachärzte einzubeziehen. Für manche Versorgungsbereiche kann es sinnvoll sein, unterschiedliche Fachdisziplinen einzubinden.

Pflegefachkräfte

Pflegekräfte spielen in der gesundheitlichen Versorgung eine zentrale Rolle und beeinflussen damit die Ergebnisse der Qualitätsindikatoren. Daher ist in der Regel mindestens eine Pflegefachkraft in jeder Fachkommission einzubeziehen.

Klinische Qualitätsmanagerin bzw. klinischer Qualitätsmanager

Ausgehend von der Qualitätsbewertung sollten Empfehlungen für geeignete Maßnahmen der Qualitätsverbesserung vorgeschlagen werden. Diese Maßnahmen sollten idealerweise unter Berücksichtigung gängiger Qualitätsmanagementstrategien erfolgen, daher sollte mindestens eine Expertin bzw. ein Experte des klinischen Qualitätsmanagements hinzugezogen werden.

Patientenvertreterinnen bzw. -vertreter

Als Empfänger der Gesundheitsleistungen, deren Qualität gesichert werden soll, ist die Perspektive der Patientinnen und Patienten im Rahmen von Qualitätssicherungsverfahren unverzichtbar. Bei allen QS-Verfahren ist daher mindestens eine Patientenvertreterin bzw. ein Patientenvertreter mit Kenntnissen im Bereich des jeweiligen Versorgungsbereichs durch Vorschlag der Patientenvertretung nach § 140f SGB V einzubeziehen.

In Abhängigkeit vom jeweiligen QS-Verfahren sollten folgende **weitere Expertinnen und Experten** benannt werden:

Weitere Angehörige der Gesundheitsfachberufe

Zu den weiteren Gesundheitsfachberufen zählen z. B. Logopädinnen und Logopäden, Ergotherapeutinnen und Ergotherapeuten, Hebammen und Entbindungshelfer, Physiotherapeutinnen und Physiotherapeuten. Die Ergebnisse der Qualitätsindikatoren werden durch die Behandlungsergebnisse der Mitglieder dieser Berufsgruppen beeinflusst (z. B. im Versorgungsbereich Orthopädie und Unfallchirurgie[73]). Daher sollte in Abhängigkeit vom jeweiligen Versorgungsbereich mindestens eine Expertin bzw. ein Experte aus einem für das QS-Verfahren relevanten Gesundheitsfachberuf einbezogen werden.

Krankenhaushygienikerin bzw. -hygieniker oder Hygienefachkraft

Bei allen Versorgungsbereichen, bei welchen die Ergebnisse der Qualitätsindikatoren durch die Krankenhaushygiene beeinflusst werden, sollte mindestens ein Krankenhaushygieniker bzw. eine Hygienefachkraft eingebunden werden.

Hinsichtlich der Einbindung bestimmter Expertinnen und Experten nach der Art der Berufsgruppe ist auch die Anzahl der jeweiligen Expertinnen und Experten abhängig vom jeweiligen Versorgungsbereich und sollte jeweils bundeseinheitlich geregelt sein.

Assoziierte Mitglieder

Für spezifische Fragestellungen können zusätzlich auch weitere Personen für die jeweilige Sitzung und nicht über ein Verfahrensjahr hinausgehend von der LAG bzw. Bundesstelle beratend hinzugezogen werden. Dies können z. B. Assistenzärztinnen und -ärzte, medizinische Dokumentarinnen und Dokumentare und andere Vertreterinnen und Vertreter von Berufsgruppen der Gesundheitsversorgung sein, aber auch Expertinnen und Experten, die sich im Ruhestand noch aktiv an der Qualitätssicherung ihres Fachgebiets beteiligen. Die in Abschnitt 6.7.1.2 genannten formalen Anforderungskriterien sind für die assoziierten Mitglieder einer Fachkommission ebenfalls zu beachten. Die Berufung der assoziierten Mitglieder erfolgt durch die LAG bzw. Bundesstelle. Die Gründe für die Berufung der Expertinnen und Experten macht die LAG bzw. Bundesstelle in geeigneter Form transparent.

Die LAG bzw. Bundesstelle trägt die Verantwortung für die Prüfung und Einhaltung sowohl der Kriterien für die Auswahl der Expertinnen und Experten als auch der Kriterien zur Zusammensetzung der Fachkommission und veröffentlicht diese für die jeweilige Fachkommission auf der Webseite der Geschäftsstelle der LAG bzw. des IQTIG. Wenn von den obigen Kriterien abgewichen werden musste (z. B. da zu wenig Fachexpertinnen bzw. Fachexperten mit entsprechender fachlicher Qualifikation zur Verfügung standen), sind die Abweichung und deren Gründe zu dokumentieren und zu veröffentlichen.

[73] Zum Beispiel Qualitätsindikatoren zur Knie- und Hüftbeweglichkeit durch Maßnahmen der Physiotherapeutinnen und Physiotherapeuten.

6.8 Fazit

Das in den vorherigen Abschnitten beschriebene Vorgehen für die fachliche Bewertung stellt eine konsequente Umsetzung des Ziels dar, möglichst explizite Verfahrensregeln zu formulieren und auf heuristische Beurteilungsprozesse nur dort zurückzugreifen, wo sie unvermeidlich erscheinen. Es wurde dargestellt, wie ein Maximum an Standardisierung und damit Objektivität der fachlichen Bewertung theoretisch erreicht werden könnte. Dabei wurde deutlich, dass die Steigerung der Objektivität des Verfahrens durch Standardisierung mittels expliziter Regeln an Grenzen stößt. Erstens lassen sich aufgrund der qualitativen Informationsgrundlage in den Stellungnahmen nicht alle Schritte der fachlichen Bewertung standardisieren. Zweitens geht eine höhere Standardisierung auch mit teilweise komplexeren Prozessen einher (vgl. Abschnitt 3.6). So müssen dafür beispielsweise die in der Stellungnahme vorgebrachten Gründe systematisch klassifiziert werden, eine Nachberechnung vorgenommen und ein Abgleich mit einem (modifizierten) Referenzwert durchgeführt werden. Der im Anhang in Kapitel 8 dargestellte Gesamtalgorithmus bei einem möglichst expliziten Vorgehen verdeutlicht diese Komplexität. Allerdings müssten die meisten der hier vorgeschlagenen, explizit formulierten Schritte auch bei einem rein heuristischen Vorgehen so oder so ähnlich angewendet werden (siehe Abschnitt 6.6). Wird beispielsweise bei der Prüfung, ob der Ausschluss bestimmter Fälle dazu führt, dass der Indikatorwert den Referenzbereich erreicht, nicht nach explizit formulierten Regeln vorgegangen, ist das Vorgehen nur schwer überprüfbar und fehleranfällig.

Wie in Abschnitt 3.6 angedeutet, sollte für die Umsetzung der Empfehlungen für die fachliche Bewertung eine Kosten-Nutzen-Abwägung durchgeführt werden. Der Nutzen eines möglichst expliziten Vorgehens in Form größtmöglicher Objektivität und Transparenz muss gegenüber den Kosten, die die Umsetzung des Vorgehens bedeuten würde, abgewogen werden. Hinweise, wo mit vermehrtem Aufwand im Vergleich zum bisherigen Vorgehen zu rechnen ist, wurden in Abschnitt 6.6 ausgeführt. Der Nutzen von Entscheidungsregeln, die explizit gemacht werden, besteht u. a. darin, dass sie auch für Kritik zugänglich werden (Dawes et al. 1989). Implizite Entscheidungsregeln dagegen können nicht überprüft und demnach auch nicht kritisiert werden. Wie in Kapitel 2 ausgeführt, werden die Ergebnisse der indikatorgestützten externen Qualitätssicherung sowohl für die Qualitätsförderung als auch für Zwecke der *accountability* eingesetzt. Dabei stellt der zweitgenannte Verwendungsbereich besonders hohe Anforderungen an die Güte der Qualitätsmessungen (Chassin et al. 2010, Solberg et al. 1997). Die Güte der fachlichen Bewertung muss sich daher an diesen höheren Anforderungen an die Objektivität und Transparenz im Verwendungszweck *accountability* orientieren. Es wird daher empfohlen, in jedem Fall den expliziten Bewertungsalgorithmus gemäß Abbildung 33 für die Einstufung der Ergebnisse in „Qualitätsdefizit“ und „Hinweis auf Qualitätsdefizit entkräftet“ umzusetzen. Für die Klassifikation in die Ziffern des Bewertungsschemas (Unterkategorien) kann dagegen statt der empfohlenen expliziten Bewertungsalgorithmen (Abschnitte 6.4.4 und 6.4.5) auch ein heuristisches Vorgehen umgesetzt werden, wenn der Gewinn an Objektivität und Transparenz gegenüber dem Aufwand für die Umsetzung weniger stark gewichtet wird. Allerdings führt der Verzicht auf eine Differenzierung hier nur zu einer geringen Aufwandsreduktion.

7 Modul Qualitätsförderung

Maßnahmen zur Verbesserung der Versorgungsqualität stellen neben der Messung und Darstellung der Versorgungsqualität die zweite wichtige Säule der gesetzlich verpflichtenden Qualitätssicherung dar. Die empfohlene Standardisierung des Mess- und Bewertungsprozesses sowie die Trennung zwischen den Modulen „Qualitätsbewertung“ und „Qualitätsförderung“ sollen dazu beitragen, dass den LAG bzw. der Bundesstelle und den Leistungserbringern mehr Ressourcen für qualitätsverbessernde Maßnahmen zu Verfügung stehen (siehe hierzu auch Kapitel 8). Das Modul „Qualitätsförderung“ beinhaltet diejenigen Maßnahmen, die nach Abschluss des Moduls „Qualitätsbewertung“ zur Verbesserung der Versorgungsqualität eingesetzt werden können. Für die Verbesserung der Versorgungsqualität werden insbesondere die dialogischen Anteile des bisherigen Strukturierten Dialoges benötigt. Hierzu zählen z. B. Besprechungen und Begehungen. Während ein qualitatives, dialogisches Vorgehen für objektive und standardisierte Qualitätsmessungen als weniger geeignet angesehen wird, soll es berechtigter Weise im Rahmen der Qualitätsförderung in besonderem Maß zum Einsatz kommen. Für eine effektive Qualitätsförderung müssen Sachverhalte erklärt und Hypothesen über geeignete Maßnahmen zur Qualitätsverbesserung entwickelt werden, wozu sich in erster Linie qualitative Methoden eignen (IQTIG 2019a: Abschnitt 5.1). Folglich wird eine Verschiebung dieser Anteile vom Modul der Qualitätsbewertung zum Modul der Qualitätsförderung empfohlen.

Die auf die Qualitätsbewertung aufbauende Verbesserung der Versorgungsqualität durch den Leistungserbringer entspricht teilweise der Verwendung von Qualitätsmessungen im Rahmen der *improvement* Strategie nach Berwick et al. (Berwick et al. 2003). Dabei wird davon ausgegangen, dass die Ergebnisse von Qualitätsmessungen Leistungserbringer in die Lage versetzen, Defizite in ihrer Versorgung zu erkennen und durch geeignete Maßnahmen selbst beheben. Diese Strategie spiegelt sich auch in den Zielen der datengestützten Qualitätssicherung wider, die auf Grundlage der Messung der Versorgungsqualität und durch Unterstützung des einrichtungsinternen Qualitätsmanagements ein höheres Qualitätsniveau der Versorgung erreichen soll (vgl. § 1 DeQS-RL). Dabei ist eine wesentliche Voraussetzung für das Einleiten geeigneter Maßnahmen die indikatorgestützte Ursachenanalyse zu den Versorgungsdefiziten, um eine Qualitätsverbesserung zu erzielen. Ein deutlicher Unterschied in dem Vorgehen nach DeQS-RL und der *improvement* Strategie nach Berwick et al. besteht jedoch in der Freiwilligkeit der qualitätsverbessernden Maßnahmen (Berwick et al. 2003). Während die *improvement* Strategie eher von der freiwilligen Veränderung der Versorgungsgestaltung durch die Leistungserbringer auf Basis von Qualitätsergebnissen ausgeht (die sowohl intrinsisch als auch extern motiviert sein kann), kann die LAG bzw. Bundesstelle Maßnahmen für einen Leistungserbringer beschließen (vgl. § 17, Abs. 5 DeQS-RL). Für eine selbstkritische Ursachenanalyse eines Leistungserbringers sollten möglichst günstige Bedingungen geschaffen werden. Diese wurden als sogenannte Erfolgsfaktoren für qualitätsverbessernde Maßnahmen beschrieben, und deren Anwendbarkeit im Rahmen der externen Qualitätssicherung wurde in Abschnitt 3.7 erörtert. Dort wurde u. a. hervorgehoben, dass eine klare Trennung von Qualitätsbewertung und Qualitätsförderung als notwendig erachtet wird, um den Erfolg qualitätsverbessernder Maßnahmen zu steigern. Durch

diese Trennung befinden sich auch die Bewertenden in einer klaren Position, die dadurch gekennzeichnet ist, dass sie in einem kollegialen Dialog mit dem Leistungserbringer geeignete Maßnahmen entwickeln können ohne gleichzeitig eine strenge Prüfung der Versorgungsqualität vornehmen zu müssen.

Da der Schwerpunkt der Beauftragung auf der Standardisierung der Qualitätsbewertung liegt, werden im Folgenden hauptsächlich konkrete Empfehlungen für die zwingende Einleitung von qualitätsverbessernden Maßnahmen ausgesprochen. Eine Ausnahme stellen die sich anschließenden Empfehlungen zum Abschluss von Zielvereinbarungen (siehe Abschnitt 7.2) dar. Das IQTIG hat dem G-BA im Jahr 2018 ein Zielvereinbarungskonzept für die bundesbezogenen Verfahren vorgelegt, das insbesondere Empfehlungen zur Rückkopplung bezüglich des Standes der Zielerreichung zwischen IQTIG und dem Unterausschuss des G-BA beinhaltet. Um auch hier eine bundesweite Vereinheitlichung der Vorgehensweise zu erreichen, wird dieses Zielvereinbarungskonzept (siehe Abschnitt 7.2) auch für die Verwendung in den landesbezogenen Verfahren empfohlen.

7.1 Empfehlungen zur Einleitung qualitätsverbessernder Maßnahmen

Maßnahmen zur Qualitätsverbesserung sollten auf eine abgeschlossene und aussagekräftige Qualitätsbewertung aufbauen. Es wird grundsätzlich davon ausgegangen, dass das interne Qualitätsmanagement der Leistungserbringer ein Qualitätsdefizit zum Anlass nimmt, Prozesse und Strukturen zu prüfen und bei Bedarf anzupassen, um das Qualitätsdefizit zu beheben. Nach Abschluss der Qualitätsbewertung wird dem internen Qualitätsmanagement bei erstmaliger Feststellung eines Qualitätsproblems zunächst die Möglichkeit eingeräumt, intern und damit eigenverantwortlich geeignete Maßnahmen einzuleiten ohne dass die LAG bzw. Bundesstelle tätig wird. Nur bei wiederholten Qualitätsmängeln soll die LAG bzw. die Bundesstelle aktiv werden, da wiederholte Qualitätsdefizite bei einem oder inhaltlich ähnlichen Indikatoren[74] darauf hindeuten, dass die leistungserbringerinternen Anstrengungen, das Qualitätsproblem zu beheben, nicht ausreichend wirksam waren. Handelt es sich daher um ein wiederholtes Qualitätsdefizit, wird empfohlen, zwingend eine oder mehrere Maßnahmen der Maßnahmenstufe 1 und/oder der Maßnahmenstufe 2 gemäß § 17 Abs. 3 der DeQS-RL im Sinne einer Qualitätsförderung durchzuführen (siehe Abbildung 38).

[74] Auf Anregung der Landesebene (siehe Anhang, Kapitel 1) sollten die LAG auch bei wiederholten Qualitätsdefiziten bei inhaltlich ähnlichen Indikatoren die Möglichkeit haben, Maßnahmen einzuleiten. Die LAG bzw. Bundesstelle entscheidet unter Einbezug der Fachkommission, ob es sich um inhaltlich ähnliche Indikatoren handelt.

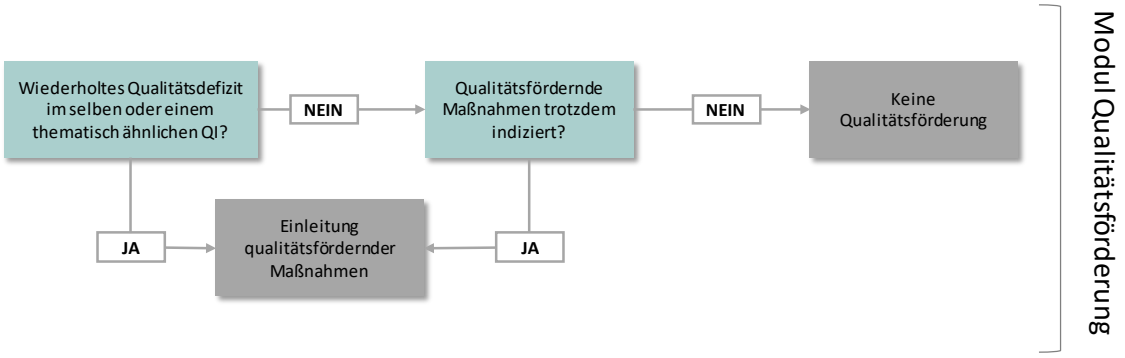

Abbildung 38: Empfehlungen zur Einleitung qualitätsverbessernder Maßnahmen (Ausschnitt aus Abbildung 9 in Kapitel 4)

Es wird daher unter Berücksichtigung der Angaben der LAG bzw. Bundesstelle zur Notwendigkeit entsprechender Maßnahmen empfohlen, dass ab dem zweiten Qualitätsdefizit bei einem Indikatorergebnis oder bei inhaltlich ähnlichen Indikatoren innerhalb von drei Jahren die LAG bzw. Bundesstelle immer qualitätsverbessernde Maßnahmen einleitet (siehe Anhang 1). Die LAG bzw. Bundesstelle kann bereits bei einem erstmalig auftretenden Qualitätsdefizit unter Angabe von Gründen jederzeit in einem Qualitätsindikator qualitätsverbessernde Maßnahmen bei einem Leistungserbringer einleiten (siehe Abbildung 38).

In § 17 Abs. 3 DeQS-RL werden qualitätsfördernde Maßnahmen in zwei Stufen unterteilt. Zu den Maßnahmen Stufe 1 zählen Vereinbarungen zwischen der LAG bzw. Bundesstelle und dem Leistungserbringer bezüglich:

- der Teilnahme an Fortbildungen, Fachgesprächen und Kolloquien,
- der Teilnahme an Qualitätszirkeln,
- der Implementierung von Behandlungspfaden,
- der Durchführung von Audits,
- der Durchführung von Peer Reviews und
- der Implementierung von Handlungsempfehlungen anhand von Leitlinien.

Nach den hier vorliegenden Empfehlungen sollen Gespräche und Begehungen zukünftig nicht mehr im Rahmen des Stellungnahmeverfahrens vorgenommen werden. Daher wird empfohlen, diese qualitätsverbessernde Maßnahmen der Maßnahmenstufe 1 zuzuordnen.

Am 18. April 2019 hat der G-BA die Richtlinie zur Förderung der Qualität und zu Folgen der Nichteinhaltung sowie zur Durchsetzung von Qualitätsanforderungen (Qualitätsförderungs- und Durchsetzungs-Richtlinie; QFD-RL) beschlossen. Diese gilt auch für die QS-Verfahren der DeQS-RL. Eine Konkretisierung anhand themenspezifischer Empfehlungen für die einzelnen QS-Verfahren soll gemäß der QFD-RL vorgenommen werden. Hinsichtlich der Maßnahmen zur Qualitätsförderung sieht die QFD-RL im Vergleich zu der Maßnahmenstufe 1 der DeQS-RL folgende weitere Beratungs- und Unterstützungsleistungen vor:

- Schriftliche Empfehlung
- Zielvereinbarungen
- Implementierung von Vorgaben für das interne Qualitätsmanagement

- Implementierung von Standard Operating Procedures (SOPs)
- Prüfung unterjähriger Auswertungsergebnisse

Beide Richtlinien sehen die Möglichkeit vor, im Anschluss an die Maßnahmen der Stufe 1 unter bestimmten Umständen weitere Maßnahmen der Stufe 2 bzw. Durchsetzungsmaßnahmen (§ 5 QFD-RL) zu veranlassen. Dazu gehören u. a. die Information zur Bewertung des Leistungserbringers und sich daraus ergebende Empfehlungen für Vergütungsabschläge des Leistungserbringers (§ 17 Abs. 4 DeQS-RL, § 5 QFD-RL)

Über die Vorgaben der beiden Richtlinien hinaus wird empfohlen, die Möglichkeit einzuräumen, qualitätsverbessernde Maßnahmen einzuleiten, wenn diese vom Leistungserbringer oder von der LAG bzw. Bundesstelle für notwendig und sinnvoll erachtet werden, um eine weitere Verbesserung der Qualität zu erreichen.

Der LAG bzw. Bundesstelle und der Fachkommission kommt damit hinsichtlich der Qualitätsförderung eine entscheidende Rolle zu. Qualitätsverbessernde Maßnahmen sollten möglichst individuell auf den einzelnen Leistungserbringer angepasst sein, um das spezifische Qualitätsproblem optimal zu adressieren. Das bedeutet, dass die LAG bzw. Bundesstelle unter Beteiligung der Fachkommission nach Erstellung einer Ursachenanalyse durch den Leistungserbringer mit diesem in Kontakt tritt, um gemeinsam geeignete Maßnahmen zu erörtern. Der Kontakt kann z. B. im Rahmen einer Besprechung oder Begehung stattfinden. Zudem trägt die LAG bzw. Bundesstelle die Verantwortung für die Umsetzung der externen Qualitätssicherung. Daher hat sie, unterstützt von der Fachkommission, neben der Bewertung der Versorgungsqualität darüber zu entscheiden, ob, wann und falls ja, welche qualitätsverbessernde Maßnahmen eingeleitet werden müssen und wie diese umzusetzen sind. Zudem wird empfohlen, den Einsatz und Umsetzungsgrad der qualitätsverbessernde Maßnahmen zu dokumentieren (siehe Kapitel 9). Die Geschäftsstelle der LAG bzw. das IQTIG hat dem Lenkungsgremium regelmäßig über den Stand der Maßnahmen zu berichten.

In Abschnitt 3.7 wurde die fachliche Kompetenz der Peers als Erfolgsfaktor für qualitätsverbessernde Maßnahmen identifiziert (siehe auch die Empfehlungen in Abschnitt 6.6). Da die fachliche Kompetenz der Peers nicht nur im Rahmen der Qualitätsbewertung für die Akzeptanz des Gesamtverfahrens von Bedeutung ist, sondern auch im Setting der Qualitätsförderung, erscheint es zukünftig von Bedeutung ein „Lernen von den Besseren“ für diesen Bereich zu etablieren. Ein erster Schritt und zugleich die Voraussetzung, um ein Lernen von den Besseren zu ermöglichen, ist die in Abschnitt 6.1 ausgesprochene Empfehlung, die Methodik dahingehend weiterzuentwickeln, dass zureichende Qualität bezogen auf das Qualitätsziel eines Indikators festgestellt werden kann. Somit bestünde grundsätzlich die Möglichkeit, diejenigen Leistungserbringer, bei denen mit hoher Sicherheit davon ausgegangen werden kann, dass sie die Qualitätsanforderungen erfüllt haben, zu identifizieren und als Peers für qualitätsverbessernde Maßnahmen bei anderen Leistungserbringern zu gewinnen. So lange das Verfahren zur Qualitätsmessung jedoch nur das Nicht-Erreichen von Qualitätsanforderungen feststellen kann, ist ein Lernen von den Besseren nur schwierig zu realisieren.

7.2 Zielvereinbarungen

Im Rahmen von Zielvereinbarungen einigen sich ein Leistungserbringer und die LAG bzw. Bundesstelle auf Empfehlungen zur Behebung von Qualitätsdefiziten (fachliches Ziel) in einem festgelegten Zeitraum (Zeitziel) sowie ggf. auf begleitende Maßnahmen, die dieser Zielerreichung dienlich sind (flankierende Umsetzungsziele).

Abgeschlossene Zielvereinbarungen weisen dem Leistungserbringer also konkrete Handlungsempfehlungen zu, die für eine Verbesserung der Versorgungsqualität von den Beteiligten als notwendig erachtet werden. Bis zur Zielerreichung wird der weitere Verlauf vom internen Qualitätsmanagement der Einrichtung beobachtet und der Erfolg der Maßnahmen dem Lenkungsgremium der LAG nach § 5 DeQS-RL berichtet.

7.2.1 Formulierung von Zielvereinbarungen

Als erster Schritt zum Abschluss von Zielvereinbarungen erfragt die LAG bzw. Bundesstelle, welche Maßnahmen die Einrichtung vorschlägt, um bestehende Qualitätsdefizite zu beheben. Auch die prüfende Fachkommission kann z. B. in einer ihrer Sitzungen Rückmeldungen geben, wo sie Möglichkeiten für Verbesserungen sieht. Diese Überlegungen finden in kollegialer Weise statt und die Vorschläge werden gesammelt und in Zielvereinbarungen zusammengetragen.

Zielvereinbarungen sollten die folgenden Hauptinhalte enthalten:

- kurze Zusammenfassung der Situation
- Aufzählung der mit einem Qualitätsdefizit bewerteten Indikatoren, einschließlich des Zeitraums, innerhalb dessen erste Besserungen eintreten sollen
- Aufzählung der von der Einrichtung weiter durchzuführenden Ursachenanalysen der mit einem Qualitätsdefizit bewerteten Indikatoren einschließlich der Berichtszeitpunkte
- Aufzählung der vereinbarten Maßnahmen zur Unterstützung der Behebung der Qualitätsdefizite einschließlich der Berichtszeitpunkte über die Durchführung der Maßnahmen und der Erfolgskontrollen
- weitere spezifische Vereinbarungen, sofern erforderlich

Folgende ergänzende Inhalte können Zielvereinbarungen darüber hinaus enthalten:

- Vereinbarung zusätzlicher Maßnahmen zur weiteren Klärung der Ursachen des Qualitätsdefizits, z. B. Bereitstellung von weiteren Materialien und/oder Begehung vor Ort
- Vereinbarung flankierender unterstützender Maßnahmen (z. B. Peer Review, Einbindung von Expertinnen und Experten)
- Vereinbarung zur Lieferung weiterer Informationen durch den Leistungserbringer, die den Fortgang der Verbesserungsmaßnahmen dokumentieren (interne Statistiken, revidierte SOPs, Bericht zu strukturellen Maßnahmen etc.)
- Vereinbarung von Maßnahmen zur Abschluss- oder Nachhaltigkeitskontrolle

7.2.2 Übersicht über den Stand der Zielvereinbarungen

Die Geschäftsstelle der LAG bzw. das IQTIG gibt dem Lenkungsgremium nach routinemäßig oder bei besonderem Anlass eine Übersicht zum Sachstand aller jeweils aktiven Zielvereinbarungen

(siehe Kapitel 9). Darüber hinaus wird zu Zielvereinbarungen, in deren Verlauf ein Qualitätsdefizit bei Indikatoren gemessen wird, gesondert berichtet.

Im Anhang findet sich ein Formular zur Standardisierung eines solchen Einzelberichts mit Hinweisen und Erläuterungen zur Nutzung (siehe Anhang, Kapitel 8).

Zu jeder seiner Sitzungen wird dem Lenkungsgremium über den aktuellen Stand von Zielvereinbarungen berichtet. Dies erfolgt in tabellarischer Form und unter Wahrung der Anonymität der Leistungserbringer.

Aufgeführt wird:

- die Anzahl der zum jeweils aktuellen Zeitpunkt aktiven Zielvereinbarungen
- die Anzahl der aktiven Zielvereinbarungen, die zum aktuellen Zeitpunkt eingehalten werden
- die Anzahl der aktiven Zielvereinbarungen, die zum aktuellen Zeitpunkt nicht eingehalten werden
  - bei Nicht-Einhaltung von mind. einer Zielvereinbarung wird dies in der Spalte „Erläuterung“ näher beschrieben
- die Anzahl der von der LAG vorgesehenen, jedoch nicht abgeschlossenen Zielvereinbarungen
  - Der Hintergrund, weshalb Zielvereinbarungen nicht abgeschlossen wurden, wird in der Spalte „Erläuterung“ näher beschrieben (z. B. Verweigerung der Unterzeichnung durch den Leistungserbringer)

Tabelle 15: Vorlage für den Bericht an das Lenkungsgremium zum aktuellen Stand aktiver Zielvereinbarungen

	Anzahl	**Erläuterung**
aktive Zielvereinbarungen		…
davon: Zielvereinbarung zum aktuellen Zeitpunkt eingehalten		…
davon: Zielvereinbarung zum aktuellen Zeitpunkt nicht eingehalten		…
geplante, nicht abgeschlossene Zielvereinbarungen		…
seit dem letzten Bericht abgeschlossene Zielvereinbarungen		…

8 Möglichkeiten zur Verkürzung des Stellungnahmeverfahrens

Die Beauftragung sieht vor, dass das IQTIG Optionen zur Verkürzung des Verfahrens[75] prüft. Gemäß der hier vorliegenden Ausführungen und Empfehlungen (siehe Kapitel 2, 4 und 7) soll das Verfahren zukünftig in die verschiedenen Module Qualitätsbewertung, Bewertung der Dokumentationsqualität und Qualitätsförderung aufgeteilt werden. Da aussagekräftige Qualitätsbewertungen die Ausgangsbasis für alle folgenden Verwendungszwecke darstellen, wurden Optionen zur Verkürzung des Verfahrens vor allem mit Blick auf das Modul Qualitätsbewertung geprüft. Erstens ist aus der Perspektive der Patientinnen und Patienten eine zeitnahe Verfügbarkeit der abschließenden Qualitätsbewertung wünschenswert, beispielsweise für Auswahlentscheidungen. Eine zeitnahe Qualitätsbewertung würde zweitens auch eine frühere Einleitung von qualitätsverbessernden Maßnahmen und mehr Zeit für entsprechende Maßnahmen ermöglichen, deren Erfolg unter Umständen dann bereits im Jahr nach der Datenerfassung (Erfassungsjahr+1) (EJ+1) quantifizierbar wäre.

8.1 Das Verfahren gemäß DeQS-RL

Da neben dem Stellungnahmeverfahren selbst auch die vor- und nachgelagerten Phasen einen Einfluss auf die Dauer des Stellungnahmeverfahren haben, werden in Abbildung 39 die Phasen des Verfahrens sowie die Fristen der DeQS-RL schematisch dargestellt. Für die Qualitätsindikatoren der Verfahren nach DeQS-RL werden zukünftig unterschiedliche Datenquellen herangezogen werden (einrichtungs- und fallbezogene QS-Dokumentation, Sozialdaten und Patientenbefragungen). Da bisher jedoch nur einrichtungs- und fallbezogenen QS-Dokumentation eingesetzt werden, beziehen sich die nachfolgenden Kapitel nur auf diese Datenquellen.

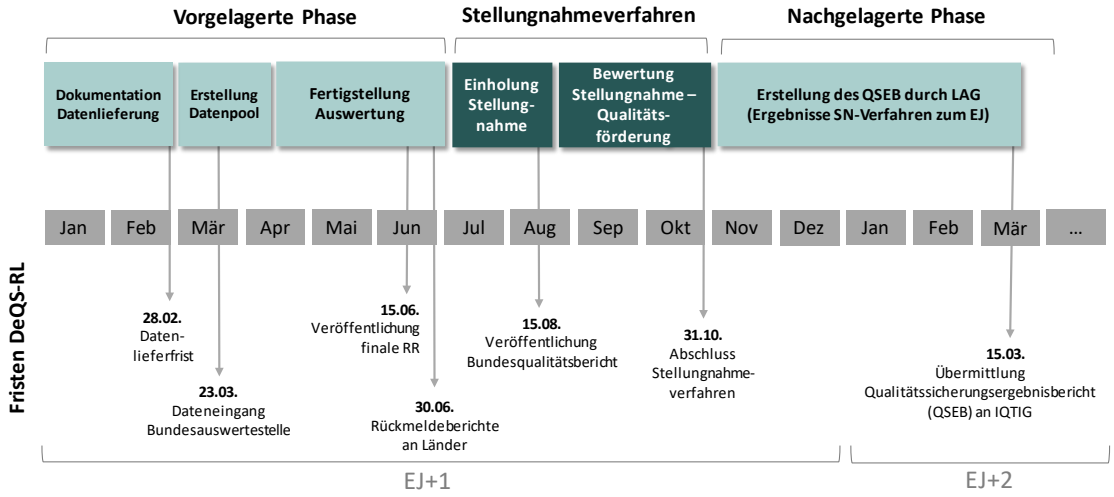

Abbildung 39: Phasen und Fristen des Verfahrens gemäß DeQS-RL.

[75] Unter Verfahren wird das Gesamtverfahren von der Dokumentation und Datenlieferung bis zur Berichterstattung verstanden.

In der vorgelagerten Phase können drei Prozesse unterschieden werden. Hierzu zählen die Dokumentation und Datenlieferung der Leistungserbringer bzw. die Datenübermittlung an die Bundesauswertungsstelle. Gemäß den themenspezifischen Bestimmungen der DeQS-RL (für *QS WI* und *QS PCI*) ist die Frist für die Datenlieferung des gesamten Erfassungsjahres (EJ) der 28. Februar des darauffolgenden Jahres (EJ + 1). Nach Ablauf der Korrekturfrist am 15. März werden die Daten von der Vertrauensstelle pseudonymisiert, sodass die Daten erst am 23. März bei der Bundesauswertestelle eingehen. Anschließend folgt die Erstellung des Bundesdatenpools durch das IQTIG. Der letzte Schritt in der vorgelagerten Phase ist die Fertigstellung der Qualitätsindikatorendatenbank (QIDB) und die Auswertungserstellung gemäß den finalen Rechenregeln (einschließlich der Entwicklung von Risikoadjustierungsmodellen anhand der Daten aus dem Erfassungsjahr) bis zum 15. Juni des EJ + 1. Der Versand der Rückmeldeberichte an die LAG erfolgt im Anschluss bis 30. Juni des EJ + 1. Parallel zu den genannten Prozessen erfolgt die Erstellung des Bundesqualitätsberichts, der am 15. August des EJ + 1 veröffentlicht wird und der neben den reinen Indikatorergebnissen des EJ auch die Ergebnisse des Stellungnahmeverfahrens zum Bezugsjahr EJ – 1 enthält (siehe Abbildung 39).

Im Stellungnahmeverfahren, welches gemäß DeQS-RL für die Krankenhäuser bis zum 31. Oktober des EJ + 1 abzuschließen ist, können zwei Prozesse unterschieden werden. Der erste Prozess bezieht sich auf das Einholen der Stellungnahmen, in welchem die Stellungnahmen angefordert, erstellt und abgegeben werden. Im sich anschließenden Prozess findet die Bewertung der Stellungnahmen sowie die Entscheidung darüber, welche qualitätsverbessernde Maßnahmen eingeleitet und umgesetzt werden sollen sowie die Umsetzung (durch die LAG bzw. das IQTIG z. B. in Form kollegialer Gespräche, Begehungen, Zielvereinbarungen etc.) selbst statt. Da analog zur QSKH-RL auch gemäß DeQS-RL die Qualitätsbewertung nicht zwingend vor den Maßnahmenstufen abgeschlossen sein muss, beinhaltet der letzte Prozess eben auch die Umsetzung der Maßnahmen der Qualitätsverbesserung.

In der nachgelagerten Phase erfolgt die Veröffentlichung der qualitativen Bewertung nach dem Stellungnahmeverfahren und der durchgeführten oder zur Durchführung empfohlenen Qualitätssicherungsmaßnahmen im Bundesqualitätsbericht bis zum 15. August. Um den Bundesqualitätsbericht erstellen zu können werden die entsprechenden Ergebnisse im Qualitätssicherungsergebnisbericht (QSEB) zusammengefasst und bis zum 15. März des EJ + 2 an das IQTIG übermittelt (siehe Abbildung 39).

8.2 Empfehlungen zur Verkürzung des Stellungnahmeverfahrens

Eine Verkürzung des Stellungnahmeverfahrens kann grundsätzlich auf drei unterschiedliche Weisen erreicht werden. Eine Möglichkeit ist, den Aufwand für die LAG bzw. Bundesstelle zu reduzieren, indem z. B. die Anzahl zu prüfender Qualitätsindikatorergebnisse verringert wird. Dies kann über die Anzahl an ausgewerteten Qualitätsindikatoren sowie über die Methodik zur Berechnung der quantitativen Auffälligkeitseinstufung für diese Qualitätsindikatoren gesteuert werden. Dadurch stünden mehr Ressourcen für die Bearbeitung der verbleibenden Indikatorergebnisse zur Verfügung und die Bearbeitung könnte unter Umständen beschleunigt werden. Eine zweite Möglichkeit besteht darin, die Prozesse des Stellungnahmeverfahrens effizienter zu

gestalten. Drittens besteht die Möglichkeit, mehr Ressourcen zur Verfügung zu stellen, z. B. indem mehr Personal bei den LAG eingesetzt wird, um die Bearbeitung zu beschleunigen. Im Folgenden werden Optionen zur Verkürzung des Verfahrens über eine Aufwandsreduktion und Effizienzsteigerung der Prozesse ausgehend von den hier vorgelegten Empfehlungen und den Rückmeldungen der Vertreterinnen und Verteter der LAG und LQS im Rahmen des Workshops zur Weiterentwicklung des Strukturierten Dialogs erörtert.

8.2.1 Aufwandsreduktion durch Volumenreduktion

Eine Reduktion des Aufwands für die LAG bzw. Bundesstelle kann einerseits erreicht werden, indem z. B. die Anzahl an Qualitätsindikatorergebnissen, für die ein Stellungnahmeverfahren einzuleiten ist, verringert wird. Diesbezüglich werden in Kapitel 5 Empfehlungen ausgesprochen, die die Anzahl an quantitativ auffälligen Ergebnissen im Vergleich zur bisher verwendeten Methodik der rechnerischen Auffälligkeit zielgerichtet reduzieren könnte. Dadurch würde sich auch der Arbeitsaufwand bei den LAG, der Bundesstelle und Leistungserbringern im Rahmen des Stellungnahmeverfahrens reduzieren. Allerdings soll zukünftig die Möglichkeit, Hinweise zu versenden, entfallen und bei allen Qualitätsindikatorergebnissen, die einen hinreichenden Hinweis für ein Qualitätsdefizit aufweisen, soll obligatorisch eine Stellungnahme eingeholt werden. Inwieweit eine Aufwandsreduktion bei spezifischerer Auslösung eines Stellungnahmeverfahrens einerseits und gleichzeitig obligatorischer Einholung einer Stellungnahme andererseits erreicht werden kann, wird beispielhaft am QS-Verfahren HEP in Abschnitt 10.4 vorgestellt.

Eine Aufwandsreduktion kann außerdem auch durch eine Verringerung der Anzahl an Qualitätsindikatoren erreicht werden. Das IQTIG sieht vor, sukzessive die bestehenden Qualitätsindikatoren anhand der Eignungskriterien (siehe Methodische Grundlagen V1.1) zu prüfen und gegebenenfalls Empfehlungen zur Aussetzung von Indikatoren auszusprechen (siehe auch Abschnitt 11.1). Im Workshop zur Einbindung der Vertreterinnen und Vertreter der LAG und LQS äußerten diese, dass viele Indikatoren aus fachlicher Sicht nicht mehr geeignet seien und daher eine Reduzierung der Anzahl dieser Indikatoren als notwendig erachtet würde (siehe Anhang, Kapitel 1).

8.2.2 Effizienzsteigerung durch Standardisierung im Vorgehen

Das vorliegende Konzept sieht im Sinne einer Standardisierung und Effizienzsteigerung ein einheitliches schriftliches Stellungnahmeverfahren zur Qualitätsbewertung vor. Eine Standardisierung des Verfahrens erfordert für die Einholung der Stellungnahmen konkrete formale und inhaltliche Anforderungen an die Stellungnahmen der Leistungserbringer (siehe Abschnitt 6.5). Es wird außerdem empfohlen, dass die LAG bzw. die Bundesstelle bei Nicht-Einhaltung der Anforderungen direkt eine abschließende Bewertung ohne weitere Einholung von Präzisierungen vornehmen soll (siehe Kapitel 6). Diese Empfehlung soll die LAG bzw. Bundesstelle einerseits dadurch entlasten, dass diese die Anonymisierung und Pseudonymisierung nicht nachträglich selbst vornehmen müssen, falls dies seitens der Leistungserbringer nicht erfolgt ist. Anderseits würde dieses Vorgehen den Aufwand der LAG bzw. Bundesstelle auch dadurch reduzieren, dass eine weitere inhaltliche Prüfung aufgrund der Einholung von Präzisierungen entfallen würde. Dabei ist zu berücksichtigen, dass auch die formale Prüfung einen gewissen Aufwand darstellt, der jedoch geringer ausfallen dürfte als eine inhaltliche Prüfung von Präzisierungen.

Neben den Empfehlungen zu Fristen im Rahmen der formalen Anforderungen an das Stellungnahmeverfahren wurden in den vorherigen Kapiteln weitere Empfehlungen zur Standardisierung des Bewertungsprozesses gemacht. Eine Standardisierung von Prozessen steigert z. B. dadurch die Effizienz, dass Entscheidungen, die zuvor von den LAG bzw. der Bundesstelle (Versand von Hinweisen) oder den Fachkommissionen getroffen wurden, zukünftig einheitlich geregelt wären. So entfiele beispielsweise durch die Einholung einer Stellungnahme bei Vorliegen eines hinreichenden Hinweises auf ein Qualitätsdefizit der Bedarf für eine Fachkommissionssitzung, da die Entscheidung über die Einholung einer Stellungnahme nach statistischen Kriterien einheitlich getroffen würde. Auf der anderen Seite geht die empfohlene Umstellung der fachlichen Bewertung weg von einem bisher eher heuristischen Vorgehen hin zu einem stärker expliziten Vorgehen einschließlich Nachberechnung des Indikatorwerts und Vergleich mit dem Referenzwert vermutlich in der Anfangsphase mit einem höheren Aufwand als bisher einher (vgl. Abschnitt 6.6).

8.2.3 Auswirkungen der Empfehlungen auf den Beginn der Bewertung von Stellungnahmen

Die hier vorgelegten Empfehlungen zu bundesweit einheitlichen Fristen für die Abgabe einer Stellungnahme von fünf Wochen würden vor dem Hintergrund des Beginns des Stellungnahmeverfahrens am 30. Juni des EJ + 1 einen klar definierten Stellungnahmezeitraum ohne Zeitverzug ermöglichen. Nach Eingang der Rückmeldeberichte bei den LAG benötigen diese Zeit, die Aufforderungsschreiben zu erstellen und zu versenden, sodass die Aufforderung zur Stellungnahme eine Woche später (am 07. Juli) beim Leistungserbringer eingehen könnten. Dieses Zeitfenster erscheint auch vor dem Hintergrund der vorliegenden Empfehlungen, die bei bestehendem hinreichendem Hinweis auf ein Qualitätsdefizit keinen Ermessensspielraum bei der Einholung einer Stellungnahme erlauben, als ausreichend. Anders als in den themenspezifischen Bestimmungen gemäß § 12 DeQS-RL festgelegt, würde damit die Einholung der Stellungnahmen nicht erst von der Fachkommission beraten werden, sondern unmittelbar durch die LAG auf Basis des statistischen Kriterien erfolgen, woraus sich eine deutliche Zeitersparnis ergeben dürfte (siehe Abschnitt 8.3). Zudem ist eine Aufwandsreduktion dadurch möglich, dass die Aufforderungen zur Stellungnahme nicht mehr individuell angepasst werden müssen, sondern einheitlich für alle QI wären (siehe Abschnitt 6.5 und Anhang, Kapitel 7). Somit müsste von den LAG ausschließlich die Auswertungen für einen Leistungserbringer zusammen mit einem einheitlichen Begleitschreiben weitergeleitet werden. Unter Berücksichtigung des Stellungnahmezeitraums von fünf Wochen und einem einwöchigen Zeitraum bis zum Eingang der Aufforderung zur Stellungnahme bei den Leistungserbringern könnte somit die Abgabefrist für Stellungnahmen bzw. die Bewertung der Stellungnahmen sechs Wochen nach dem 30.06. des EJ+1 also auf den 11. August beginnen.

8.2.4 Auswirkungen der Empfehlungen auf den Abschluss des Stellungnahmeverfahrens

Eine frühere Verfügbarkeit der Qualitätsergebnisse kann auch über die in Kapitel 4 beschriebene konzeptuelle und chronologische Trennung von Qualitätsbewertung und -förderung erreicht werden. Diese Trennung würde neben einer Effektivitätssteigerung der Qualitätsförderung

(siehe Kapitel 4) auch den Aufwand im bisherigen Stellungnahmeverfahrens reduzieren, da der Prozess der Qualitätsförderung mit den Leistungserbringern in der nachgelagerte Phase zu verorten wäre.

Derzeit beträgt der Zeitraum für das gesamte Stellungnahmeverfahren ab Versand der Rückmeldeberichte an die LAG bis Abschluss des Stellungnahmeverfahrens gemäß DeQS-RL vom 30. Juni bis 31. Oktober, also vier Monate (siehe Abbildung 39). Unbekannt ist, innerhalb welches Zeitrahmens die LAG derzeit die Bewertung der Stellungnahmen einschließlich der Aufbereitung der Stellungnahmen vornehmen. Es ist jedoch bekannt, dass neben dem Zeitraum für die Bewertung der Stellungnahmen innerhalb des 4-monatigen Zeitraums außerdem folgende Aufgaben anfallen:

1. die Entscheidung, von welchen Leistungserbringern Stellungnahmen eingeholt werden sollen
2. die Erstellung und der Versand der Aufforderungsschreiben für die Stellungnahmen
3. der Stellungnahmezeitraum
4. mögliche Präzisierungen

Für 1. kann unter der Berücksichtigung, dass die Fachkommission gemäß DeQS-RL einzubeziehen ist (siehe Abbildung 40), vermutlich eine Woche geschätzt werden. Für 2. muss zum Eingang der Aufforderung zur Stellungnahme beim Leistungserbringer vermutlich mindestens eine weitere Woche geschätzt werden. Dieses Zeitintervall ist sicherlich stark von dem Grad der Automatisierung dieses Prozesses bei den LAG abhängig. Hinsichtlich 3. wurden im Workshop Zeiträume von vier oder fünf Wochen, als umsetzbar diskutiert (siehe Anhang, Kapitel 1). Für das Einholen und das Verfassen von Präzisierungen zu Stellungnahmen können weitere ein bis zwei Wochen geschätzt werden. Werden die geschätzten Zeitangaben von dem 4-monatigem Zeitraum subtrahiert, bedeutet das, dass für die Bewertung der Stellungnahmen und Präzisierungen (inkl. Aufbereitung der Stellungnahmen durch die LAG bzw. Bundesstelle) ca. acht Wochen zur Verfügung stehen.

Ausgehend von einem Start der Bewertung der Stellungnahmen am 11. August, der Verlagerung der Qualitätsförderung in die nachgelagerte Phase sowie des reduzierten Bearbeitungsbedarfs durch den Wegfall der Einholung und inhaltlichen Prüfung von Präzisierungen (siehe Abschnitt 8.2.2) erscheint eine Verkürzung der Dauer für die Bewertung der Stellungnahmen um einen Monat zunächst möglich. Folglich könnte das Stellungnahmeverfahren grundsätzlich am 30. September des EJ + 1 abgeschlossen werden (siehe Abbildung 42). Wie oben beschrieben steht diese Schätzung jedoch unter Vorbehalt, da keine Erfahrungen mit der Umsetzung der empfohlenen Methodik vorliegen. Da gerade in der Anfangsphase der Konzepteinführung jedoch mit einem höheren Aufwand zu rechnen ist, sollte der Zeitraum für die Bewertung der Stellungnahmen zunächst nicht verkürzt werden. Eine Verkürzung des Stellungnahmeverfahrens von vier auf drei Monaten sollte nur unter der Voraussetzung in Erwägung gezogen werden, dass Erfahrungen mit dem neuen Verfahren vorliegen, dieses mit einer entsprechenden IT-Unterstützung umgesetzt ist und das Stellungnahmeverfahrens vor der Sommerpause startet (siehe hierzu auch Abschnitt 8.4).

Das Modul Qualitätsförderung könnte sich direkt daran anschließen und parallel zur Berichterstattung für die Qualitätsbewertung erfolgen. Durch die Aufteilung der Prozesse in zwei Module stünde den LAG bzw. der Bundesstelle im Jahr nach der Erfassung somit mehr Zeit für die Einleitung und Umsetzung von Maßnahmen zur Qualitätsverbesserung zur Verfügung. Für den Gesamtprozess bestehend aus den Modulen der Qualitätsbewertung und der Qualitätsförderung stünden somit insgesamt sechs Monate zur Verfügung, also zwei Monate mehr als aktuell durch die DeQS-RL vorgegeben (vgl. Abbildung 39 und Abbildung 42). Zudem könnte der Erfolg mancher Maßnahmen aus dem Vorjahr ggf. bereits im EJ + 1 gemessen werden.

8.3 Vorschläge der Vertreterinnen und Vertreter der LAG und der LQS

Im Rahmen des Workshops zur Einbeziehung der Vertreterinnen und Vertreter der LAG und der LQS wurden zwei weitere Aspekte genannt, die eine Verkürzung des Verfahrens erreichen können.

Der erste Aspekt bezieht sich auf die aktuellen Entscheidungsprozesse und teilweise ungeklärten Zuständigkeiten innerhalb der LAG. Abbildung 40 zeigt den Ablauf der Bewertung von Auffälligkeiten gemäß DeQS-RL.

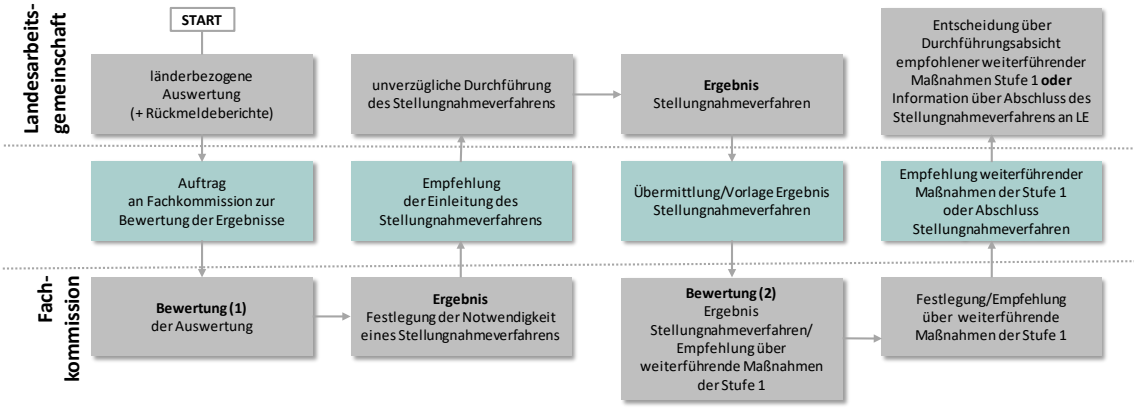

Abbildung 40: Schematischer Ablauf der Bewertung von Auffälligkeiten gemäß DeQS-RL.

Die DeQS-RL sieht aktuell vor, dass die Landesarbeitsgemeinschaften in einem Prozess wechselder Zuständigkeiten mit der Fachkommission die Bewertung der auffälligen Indikatorergebnisse vornehmen und Empfehlungen für qualitätsverbessernde Maßnahmen abgeben (Abbildung 40): Die LAG beauftragen die Fachkommissionen damit, die Ergebnisse der Qualitätsindikatoren zu bewerten. Nach dieser Bewertung durch die Fachkommission soll diese die Notwendigkeit eines Stellungnahmeverfahrens festlegen und Empfehlungen zur Einleitung des Stellungnahmeverfahrens sowie zur Art und Weise (schriftliche Stellungnahme, Gespräch oder Begehung) einschließlich des Zeitrahmens aussprechen. Die LAG sollen das Stellungnahmeverfahren durchführen und der Fachkommission das Ergebnis vorlegen, sodass diese die Ergebnisse bewerten kann. Im Weiteren empfiehlt die Fachkommission der LAG im Anschluss weiterführende Maßnahmen der Maßnahmenstufe 1 oder den Abschluss des Stellungnahmeverfahrens. Abschließend entscheidet die LAG darüber, weiterführende Maßnahmen der Maßnahmenstufe 1 durchzuführen oder gibt dem Leistungserbringer die Information, dass das Stellungnahmeverfahren abgeschlossen

ist. Die Umsetzung dieses Prozesses in der Praxis wurde im Workshop als sehr zeitaufwendig beschrieben (siehe Anhang, Kapitel 1).

Im Einklang mit dieser Rückmeldung der Vertreterinnen und Vertreter der LAG und LQS empfiehlt das IQTIG eine Vereinfachung dieses Prozesses. Zum einen entfällt der erste Entscheidungsschritt, da bei einem hinreichenden Hinweis auf ein Qualitätsdefizit immer eine Stellungnahme einzuholen ist und dies mittels der statistischen Methodik ohne Ermessensspielraum eindeutig operationalisiert ist (siehe Abschnitt 6.1). Zum anderen übernähme die LAG bzw. Bundesstelle die vorgeschaltete formale Prüfung. Damit erscheint eine vergleichsweise zeitnahe und zügige fachliche Bewertung der Stellungnahmen mit anschließender Empfehlung für die abschließende Einstufung in das Bewertungsschema durch die jeweilige Fachkommission möglich (siehe Abschnitte 6.4 und 8.2.3). Dieser auf Basis der vorliegenden Empfehlungen vereinfachte Ablauf wird in Abbildung 41 dargestellt. Hier wird auch deutlich, dass nach der formalen Prüfung und der Sitzung der Fachkommission direkt qualitätsverbessernde Maßnahmen durch die LAG bzw. Bundesstelle eingeleitet werden könnten. Welche Zuständigkeiten bei der Durchführung der Qualitätsbewertung und der Durchführung der qualitätsverbessernde Maßnahmen empfohlen werden, wird in den Kapitel 6 und 7 erläutert.

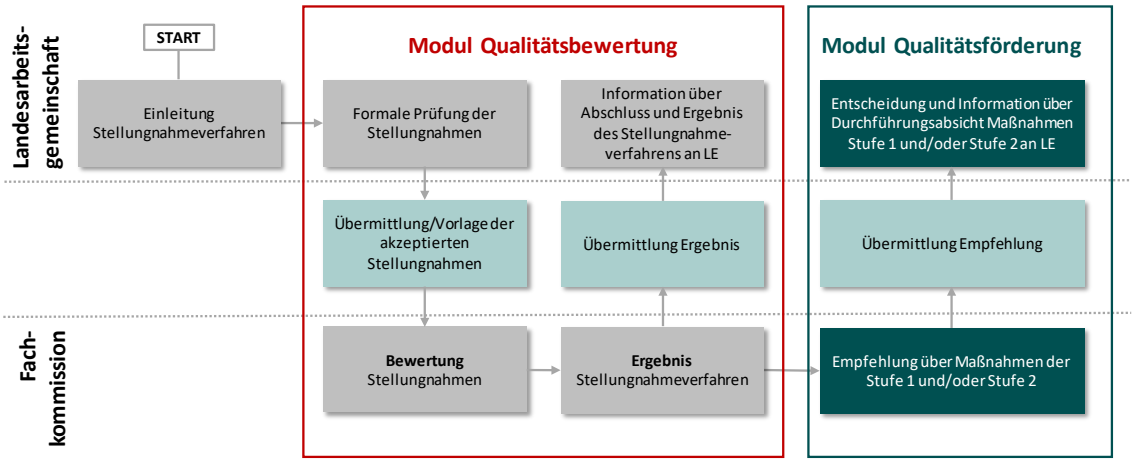

Abbildung 41: Schematischer Ablauf der Bewertung von Auffälligkeiten gemäß Empfehlungen.

Der zweite Aspekt, den die LAG und LQS hinsichtlich der Verkürzung des Verfahrens nannten, ist ein früherer Start des Stellungnahmeverfahrens (siehe Anhang, Kapitel 1). Hintergrund ist, dass aufgrund der aktuellen Fristen der DeQS-RL das Stellungnahmeverfahren in die Sommermonate fällt. Sowohl die LAG als auch die Leistungserbringer wären durch eine Vorverlagerung des Stellungnahmeverfahrens mit mehr personellen Ressourcen ausgestattet, wenn dieser Prozess nicht während der Sommerurlaubszeit mit reduziertem Personal erfolgen müsste (siehe Anhang, Kapitel 1).

8.4 Verkürzung des Gesamtverfahrens durch früheren Start des Stellungnahmeverfahrens

Eine Möglichkeit für einen früheren Start des Stellungnahmeverfahrens wäre, die Korrekturfrist (vgl. Abschnitt 8.1) für die Daten zu streichen. Angesichts der quartalsweisen Auswertungen und Rückmeldeberichte und der damit verbundenen Möglichkeit, nötige Korrekturen rechtzeitig und bereits während des EJ durchzuführen, erscheint eine zusätzliche Korrekturfrist nicht notwendig. Darüber hinaus bestünde eine weitere Möglichkeit das Stellungnahmeverfahren vorzuverlegen darin, dass man die Datenlieferfrist vorzieht. Gemäß der DeQS-RL berücksichtigt der Datensatz des EJ nur Daten der Patientinnen und Patienten, die noch im gleichen EJ entlassen werden, „Überlieger“ werden dem EJ nicht mehr zugerechnet. Dadurch ist die Dokumentation der letzten Fälle bereits am 31. Dezember des EJ möglich, sodass prinzipiell eine frühere Datenannahmefrist und damit auch ein früherer Eingang der Daten bei der Bundesauswertestelle beispielsweise bis zum 15. Februar des EJ + 1 festgelegt werden könnte. In der Folge könnte zukünftig auch der Versand der Rückmeldeberichte an die Länder früher stattfinden.

Ein früherer Start des Stellungnahmeverfahrens ist auch angesichts der Überführung der Verfahren der QSKH-RL in die DeQS-RL theoretisch möglich. Die DeQS-RL legt in § 10 Abs. 2 fest, dass das IQTIG die Auswertungen und Überprüfung der Daten auch für alle länderbezogenen Verfahren vornehmen soll. Neben einer Aufwandsreduktion für die LAG durch den Wegfall der Indikatorberechnungen (innerhalb des Prozesses „Fertigstellung Auswertung“) ergäbe sich auch eine Zeitersparnis insgesamt, da die einzelnen Prozessschritte gleichzeitig stattfinden könnten. Die Aufforderungsschreiben für die auffälligen Indikatorergebnisse könnten quasi zeitgleich mit der Übermittlung der Auswertungen von der LAG an die Leistungserbringer versendet werden (siehe Anhang, Kapitel 7), falls entsprechende IT-seitige Strukturen aufgebaut werden würden. Das bedeutet, dass mit Erstellung und Versand der Auswertung an die LAG diese direkt mit dem Stellungnahmeverfahren beginnen könnten. Demnach könnte das Stellungnahmeverfahren, welches mit der Aufforderung einer Stellungnahme bei einem hinreichenden Hinweis auf ein Qualitätsdefizit beginnt, voraussichtlich bereits mit Übermittlung der Rückmeldeberichte erfolgen. Folglich könnte auch der Start der Bewertung der Stellungnahmen eine Woche früher, am 04. August anstatt am 11. August beginnen (vgl. Abschnitt 8.2.3).

In Abbildung 42 sind die möglichen Änderungen für das Modul Qualitätsbewertung und damit die Verfügbarkeit der abschließenden Qualitätsergebnisse dargestellt.

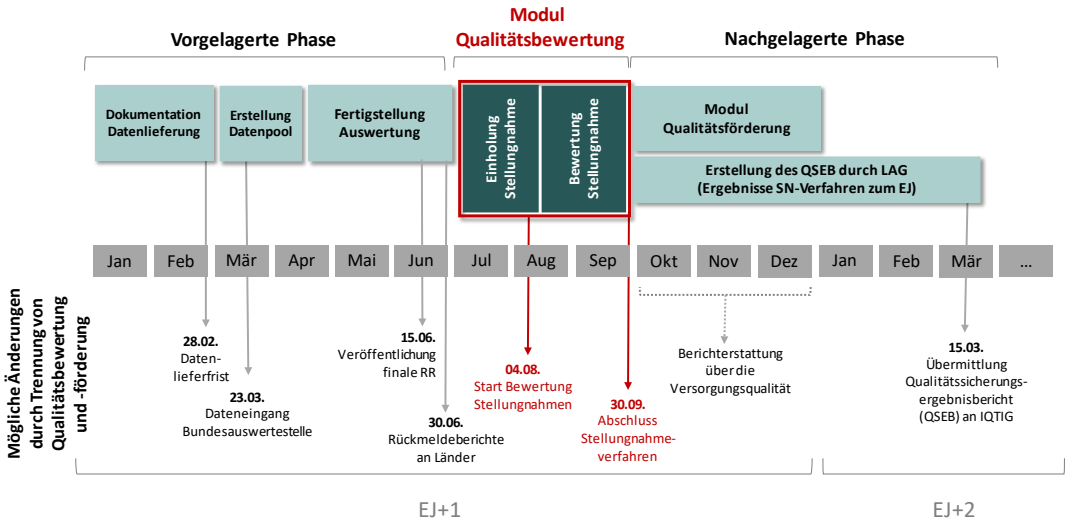

Abbildung 42: Mögliche Phasen und Änderungen des Verfahrens nach Erfahrung mit der Umsetzung der Empfehlungen.

In Kombination mit einem früheren Dateneingang bei der Bundesauswertungsstelle und damit einer möglichen früheren Übermittlung der Rückmeldeberichte bestünde grundsätzlich die Möglichkeit das Stellungnahmeverfahrens noch vor der Sommerpause beginnen zu lassen. Ein früherer Beginn des Moduls Qualitätsbewertung würde sich, wie oben beschrieben, einerseits günstig auf das Stellungnahmeverfahren auswirken und andererseits eine frühzeitigere Berichterstattung zur Versorgungsqualität ermöglichen. Eine notwendige Voraussetzung für die Festlegung entsprechender Regelungen ist jedoch eine intensive Abstimmung unter Beteiligung der unterschiedlichen Akteure (z. B. LAG, Leistungserbringer, IQTIG, Softwareanbieter, usw.). Die Abstimmung müsste beispielsweise die unterschiedlichen Abläufe der verschiedenen Verfahren (z. B. landes- und bundesbezogene Verfahren sowie plan. QI-RL-Verfahren) aber auch für die unterschiedlichen Erfassungsinstrumente (fall- und einrichtungsbezogene QS-Dokumentation, Sozialdaten und Patientenbefragung) berücksichtigen. Da ein Start des Moduls Qualitätsbewertung vor 30. Juni des EJ + 1 weitreichende Änderungen für alle Beteiligten bedeuten würde, sollten entsprechende Änderungen nicht ohne weitere umfangreiche Prüfungen der Umsetzbarkeit und der Konsequenzen vorgenommen werden (siehe hierzu auch Abschnitt 11.5.2).

In der Zusammenschau ergeben sich eine Reihe von Möglichkeiten, um einerseits das Gesamtverfahren und andererseits vor allem das Modul Qualitätsbewertung früher abzuschließen. Mit Blick auf das gesamte Verfahren entstehen im bisherigen Vorgehen Verzögerungen vor allem durch viele und teilweise längere Korrekturfristen (für Daten und für Stellungnahmen). Mit Blick auf das Modul Qualitätsbewertung sind die vorliegenden Empfehlungen zur Standardisierung der Abläufe geeignet, einige Prozesse zu automatisieren (z. B. die Entscheidung, für welchen Leistungserbringer eine Stellungnahme eingeholt werden soll) und komplexe Abstimmungsprozess zwischen Fachkommission und der LAG bzw. Bundesstelle zu vereinfachen.

9 Berichterstattung

Neben der Vereinheitlichung der Vorgehensweise im Strukturierten Dialog, der Optimierung der Effizienz des Verfahrens und der Entwicklung eines Rahmenkonzepts beinhaltet die Beauftragung auch, Empfehlungen für eine optimierte Berichterstattung zu den Ergebnissen des Strukturierten Dialogs zu erarbeiten. Dies beinhaltet Vorschläge, wie eine einheitliche Datengrundlage für den Bericht zum Strukturierten Dialog, den das IQTIG auf Basis der Berichte der Stellen auf Landesebene erstellt, gewährleistet werden kann (Punkt 2.a der Beauftragung) sowie Mindestanforderungen, die an diese Berichte der Landesstellen (Punkt 2.c der Beauftragung) gestellt sind. Während im Bericht zu Stufe 1 der Beauftragung noch Empfehlungen für die Berichte im Rahmen der QSKH-RL gemacht wurden, wird wegen des absehbaren Außerkrafttretens der QSKH-RL davon abgesehen, diese Empfehlungen weiterzuentwickeln. Daher werden an dieser Stelle Empfehlungen mit Blick auf die DeQS-RL gegeben. Somit ist ein Vorschlag für eine Vereinheitlichung der bisher bundesweit nicht einheitlich standardisierten Berichte über die Ergebnisse des Strukturierten Dialogs sowie des Datenvalidierungsverfahrens (Erstellung von Mustertabellen, Templates, wie unter Punkt 2.c der Beauftragung genannt) nicht mehr erforderlich, da sich die Inhalte dieser Berichte mit Überführung aller Verfahren in die DeQS-RL dann in den Empfehlungen zu den QSEB wiederfinden. Vor dem Hintergrund der Regelungen in der DeQS-RL, welche vorsehen, dass das IQTIG die Auswertungen für die Landesebene erstellt (§ 10 DeQS-RL), erübrigt sich auch die beauftragte Gewährleistung einer einheitlichen Datengrundlage für den Qualitätsreport und den Bericht zum Strukturierten Dialog.

Die DeQS-RL sieht vor, dass in den sog. Qualitätssicherungsergebnisberichten (QSEB) die LAG die Ergebnisse der Stellungnahmeverfahren der QS-Verfahren der externen stationären Qualitätssicherung an das Institut nach § 137a SGB V in maschinenlesbarer und -verwertbarer Form übermitteln (§19 DeQS-RL). Diese Berichte sind bis 15. März des EJ + 2 abzugeben und lösen die Berichte über die Ergebnisse des Strukturierten Dialogs sowie des Datenvalidierungsverfahrens der QSKH-RL ab. Der Bundesqualitätsbericht (BQB) nach § 20 DeQS-RL umfasst die Ergebnisse der Bundesauswertung, die Ergebnisse der Datenvalidierung, die Ergebnisse aus den QSEB sowie die Ergebnisse einer wissenschaftlichen Begleitung der Verfahren. Die bisherigen sog. „Mai-Berichte“, in denen das IQTIG die Ergebnisse der Datenvalidierung und der Strukturierten Dialoge zusammengefasst hat, werden somit zukünftig im BQB berichtet. Der Bundesqualitätsbericht ist standardisiert und wird dem G-BA in maschinenlesbarer und -verwertbarer Form übermittelt (aktuell am 15. August zu den Verfahren *QS PCI* und *QS WI*) (siehe Kapitel 8, Abb. 1)

In Kapitel 8 dieses Berichts werden Empfehlungen zur Verfahrensverkürzung gegeben. Werden diese Empfehlungen berücksichtigt, ergeben sich daraus einerseits neue Möglichkeiten für die Berichterstattung und anderseits Implikationen für die zugehörigen Datenflüsse.

Bisher liegen die Informationen aus dem Stellungnahmeverfahren, zu qualitätsverbessernden Maßnahmen und zur Datenvalidierung erst im EJ + 2 vor. Die in Kapitel 1 und Kapitel 8 vorgelegten Empfehlungen, die eine Trennung der Qualitätsbewertung und der Qualitätsförderung vor-

sehen, ermöglichen eine frühzeitigere Berichterstattung zu den Ergebnissen der Qualitätsbewertung im EJ + 1 mit einer nachgelagerten Berichterstattung zur Qualitätsförderung im EJ + 2. Es wird daher empfohlen, die Ergebnisse der fachlichen Bewertung inkl. der Begründung der Einstufung getrennt von den Empfehlungen zur Qualitätsförderung (einzuleitende Maßnahmen und deren Ergebnisse) zu berichten. Diese Trennung ermöglicht, dass die Qualitätsergebnisse früher zur Verfügung gestellt werden können, wohingegen für die Ursachenanalyse und die Durchführung qualitätsverbessernder Maßnahmen weiterhin mehr Zeit zur Verfügung stehen soll.

Es wird daher empfohlen, dass die Berichterstattung zukünftig aus den folgenden beiden „Berichtsarten" besteht, die beide zu unterschiedlichen Zeiten in den BQB einfließen:

- Berichterstattung zur Versorgungsqualität: Qualitätsergebnisse am Ende des Moduls „Qualitätsbewertung"
- Berichterstattung über die Maßnahmen der Qualitätsverbesserung: Empfehlungen für die einzuleitenden qualitätsverbessernden Maßnahmen und die Ergebnisse aus der Umsetzung dieser Maßnahmen

Für beide Berichtsarten wird eine Spezifikation dessen, was genau berichtet werden soll, benötigt. Für die Verfahren gemäß DeQS-RL liegt bereits eine Spezifikation für die Qualitätssicherungsergebnisberichte (QSEB) vor. Die Spezifikation für die QSEB kann als Grundlage für die zukünftig einheitliche Berichterstattung für alle Verfahren herangezogen werden und entspricht der im Auftrag genannten Festlegung von Mindestanforderungen für die Länderberichte zum Strukturierten Dialog inkl. der Mustertabellen (Punkt 2.c der Beauftragung). Die bisherige Spezifikation müsste jedoch an die im hier vorliegenden Konzept zur Qualitätsbewertung und -förderung vorgeschlagene Vorgehensweise angepasst werden. Darüber hinaus müssen bisher nicht abgefragte Informationen zu Maßnahmen der Qualitätsverbesserung ergänzt werden. Die konkrete Ausarbeitung kann erst nach dem Beschluss über die Umsetzung des hier vorgeschlagenen Konzepts erfolgen, da die Erstellung der Spezifikation ein aufwändiger Prozess ist, der zudem erst nach Beschluss über die Inhalte sinnvoll ist.

Bisher werden die Ergebnisse zur Versorgungsqualität und zu den qualitätsverbessernden Maßnahmen im EJ + 2 berichtet. Wünschenswert ist jedoch, dass die Ergebnisse der Qualitätsbewertung aus dem Stellungnahmeverfahren (Ende Modul Qualitätsbewertung) schon früher vorliegen. Je nach angestrebtem Erscheinungstermin für den Bundesqualitätsbericht wäre es denkbar, nach Abschluss des Stellungnahmeverfahrens die Ergebnisse der Qualitätsbewertung im Bundesqualitätsbericht zu berichten. Hierfür müsste zeitnah eine eigene Datenübermittlung der Ergebnisse des Stellungnahmeverfahrens an das IQTIG erfolgen. Für die Prüfung, Aufbereitung, Auswertung und Verschriftlichung der Ergebnisse der Qualitätsbewertung für den BQB müssen jedoch mindestens 6 bis 8 Wochen Zeit nach Abschluss des Stellungnahmeverfahrens eingeplant werden. Die Maßnahmen zur Qualitätsverbesserung sowie die Ergebnisse der Datenvalidierung würden wie bisher im EJ + 2 berichtet werden.

Sollte das hier vorgelegte Konzept zur Qualitätsbewertung und -förderung beschlossen werden, wird diesbezüglich eine Anpassung der DeQS-RL erforderlich sein (z. B. hinsichtlich der Etablierung eines Datenflusses von den LAG zum IQTIG für die Ergebnisse am Ende des Moduls Qualitätsbewertung mit der Festlegung von Fristen).

9.1 Berichterstattung zur Versorgungsqualität

Wie bereits im vorangegangenen Kapitel dargelegt, lassen sich aus den Empfehlungen zum Vorgehen bei der Qualitätsbewertung im Stellungnahmeverfahren (siehe Kapitel 6) die Vorgaben für die Spezifikation ableiten. Es ist davon auszugehen, dass Änderungen der bisherigen Spezifikation zu den in den Verfahren gemäß DeQS-RL zu übermittelnden Informationen notwendig sind. Es wird eine Leistungserbringer- und QI-bezogene Übermittlung der Ergebnisse des Stellungnahmeverfahrens empfohlen, d. h. es müssen Datensätze übermittelt werden, die für jeden Leistungserbringer und jeden QI eine Angabe beinhalten. Insbesondere müssen hier auch die Begründungen für die fachlichen Bewertungen mitgeteilt werden. Eine Aggregation der Daten (z. B. übergeordnet für das jeweilige Bundesland oder als Ländervergleich etc.) sowie eine Aus- und Bewertung erfolgt dann durch das IQTIG für den Bundesqualitätsbericht.

9.2 Berichterstattung über die Maßnahmen der Qualitätsförderung

Neben der abschließenden Bewertung der Indikatorergebnisse sind lt. DeQS-RL in den QSEB die folgenden Inhalte zu Maßnahmen der Qualitätsverbesserung zu berichten:

- die Art, Häufigkeit und Ergebnisse aller durchgeführten oder zur Durchführung empfohlenen Qualitätssicherungsmaßnahmen
- die längsschnittliche verlaufsbezogene Darstellung des Erfolgs von Qualitätssicherungsmaßnahmen
- Erfahrungsberichte aus den Fachkommissionen
- Angaben über wiederholte oder besonders schwerwiegende Auffälligkeiten

Hinsichtlich der Maßnahmen, die zur Qualitätsförderung zu ergreifen sind, finden sich Regelungen sowohl in der DeQS-RL selbst als auch übergreifend in der Richtlinie zur Förderung der Qualität und zu Folgen der Nichteinhaltung sowie zur Durchsetzung von Qualitätsanforderungen des Gemeinsamen Bundesausschusses gemäß § 137 Abs. 1 SGB V (Qualitätsförderungs- und Durchsetzungs-Richtlinie; QFD-RL). In der DeQS-RL werden zwei Maßnahmenstufen unterschieden. Maßnahmenstufe 1 enthält beispielhaft Maßnahmen, die in § 4 QFD-RL als „Maßnahmen zur Beratung und Unterstützung“ festgelegt sind; Maßnahmenstufe 2 nach DeQS-RL entspricht den „Durchsetzungsmaßnahmen“ nach § 5 QFD-RL. Wegen der übergeordneten Bedeutung der QFD-RL wird empfohlen, die dort festgelegten Maßnahmen als Grundlage für die Berichterstattung vorzusehen (zu den Maßnahmen der Qualitätsförderung siehe auch Kapitel 7).

Die Spezifikation für die Übermittlung der relevanten Informationen zu qualitätsfördernden Maßnahmen ist teilweise bereits in der Spezifikation für den QSEB umgesetzt. Dort werden Maßnahmen der Stufe 1 bzw. nach QFD-RL § 4 (Beratung und Unterstützung) abgefragt. Abfragen zu § 5 (Durchsetzung), die den Maßnahmen der Stufe 2 nach DeQS-RL entsprechen, fehlen noch.

Neben den eingeleiteten Maßnahmen selbst soll auch deren Umsetzung und das Umsetzungsergebnis berichtet werden. Dazu gehört auch ein Monitoring der Zielvereinbarungen (siehe Abschnitt 7.2). Es wird empfohlen, nach entsprechender Beschlussfassung die Informationen wie in Abschnitt 7.2 dargelegt zu berichten.

Im G-BA-Auftrag sind unter Punkt 2 „Optimierung der Transparenz und Nachvollziehbarkeit der Entscheidungsfindung“ auch ergänzende Punkte für die Berichterstattung genannt, die die Bereitstellung von Informationen für die Öffentlichkeit und für teilnehmende Krankenhäuser umfassen (Punkt 2.d). Im Rahmen des aktuellen Konzepts kann dies die Berichterstattung zu den Ergebnissen sowohl des Moduls Qualitätsbewertung als auch zum Modul Qualitätsförderung betreffen. Hinsichtlich der Aufarbeitung von Informationen für die Öffentlichkeit wird auf den Auftrag zum G-BA-Qualitätsportal verwiesen, in dem die Grundlage für eine allgemeinverständliche und für Patientinnen und Patienten nutzbringende Darstellung der Qualitätsergebnisse (Ergebnisse des Moduls Qualitätsbewertung) erarbeitet wird. Im Bundesqualitätsbericht wird die Öffentlichkeit über die Ergebnisse aus beiden Modulen informiert.

Darüber hinaus sollen zukünftig den Leistungserbringern ein gegenseitiges Lernen aus den Analysen auffälliger Ergebnisse, identifizierter Verbesserungspotenziale und beispielhaft positiver Versorgungsmodelle ermöglicht werden (Punkt 2.d und 2.e der Beauftragung, siehe auch Kapitel 7, „Lernen von den Besten“).

Der Schwerpunkt des vorliegenden Rahmenkonzepts liegt jedoch auf dem Modul „Qualitätsbewertung“. Wie oben dargelegt muss das Modul „Qualitätsförderung“ für die Berichterstattung noch im Detail spezifiziert werden. Bis die Rückmeldungen der Einrichtungen zu den mehrstufigen Maßnahmen und Ergebnissen der Qualitätsförderung tatsächlich vorliegen, werden mehrere Jahre vergehen. Daher erscheint es nicht sinnvoll, bereits in diesem Bericht und ohne einheitliche Datengrundlage Vorschläge im Detail zu entwickeln, wie diese Analyseergebnisse und Verbesserungsmaßnahmen auf einer übergeordneten Ebene allen beteiligten Einrichtungen zur Verfügung gestellt und für Fortbildungszwecke genutzt werden können. Es sollten erst die Rückmeldungen der LAG zu den getroffenen Maßnahmen und deren Erfolg etabliert sein, bevor ein Konzept für weiterführende Auswertungen dieser Daten erstellt wird.

An dieser Stelle können daher nur Eckpunkte für die zukünftige Berichterstattung zur Versorgungsqualität und zu Maßnahmen der Qualitätsverbesserung skizziert werden. Ein umfassendes Konzept mit Spezifikation für die zu übermittelnden Daten/Inhalte und zu den Zeitpunkten der Veröffentlichung muss erarbeitet werden, nachdem das vorliegende Rahmenkonzept zur Qualitätsbewertung und -förderung beschlossen wurde. Die Umsetzung dieses Rahmenkonzepts zur Qualitätsbewertung und -förderung und der dazugehörigen Berichterstattung erfordert im Vorfeld insbesondere ein schlüssiges, zeitlich machbares Ineinandergreifen aller Verfahrensabschnitte der DeQS-Verfahren und ggf. auch eine Änderung der DeQS-RL zu den dann einzuhaltenden Fristen. Dabei müssen auch Aufwand und Machbarkeit bedacht werden für die Zeit nach der Überführung aller Verfahren in die DeQS-RL. Die AG DeQS hat im Hinblick auf eine Neugestaltung des Fristengerüsts das IQTIG gebeten, die die Verfahrensabläufe der DeQS-Verfahren zu prüfen und zu überarbeiten. Die Ergebnisse dazu müssen auch im Hinblick auf die Fristen und möglichen Inhalte der Berichterstattung einbezogen und berücksichtigt werden. Die Vorschläge

zur Neugestaltung des Fristengerüsts werden aktuell parallel zum vorliegenden Bericht erarbeitet und können derzeit noch nicht in den hier vorgelegten Abschlussbericht aufgenommen werden.

10 Empfehlungen zur Evaluation des neuen Konzepts

Laut Beauftragung sollen auch Empfehlungen zur Evaluation des neuen Konzepts gegeben werden. Im folgenden Abschnitt werden Empfehlungen hinsichtlich möglicher Schwerpunkte und der Vorgehensweise einer Evaluierung des vorliegenden Konzepts dargelegt. Als Schwerpunkte der Evaluation bieten sich hierbei einerseits die Erfassung des Umsetzungsgrads und andererseits die Untersuchung der Zielerreichung an. Als Ziele der Weiterentwicklung des Strukturierten Dialogs wurden definiert:

- die Einheitlichkeit der Vorgehensweise
- die Transparenz und Nachvollziehbarkeit der Entscheidungsfindung
- die Effizienz im Sinne einer Verschlankung des Verfahrens

Eine genaue Definition der Evaluationsschwerpunkte wäre erforderlich und sollte im Hinblick auf eine mögliche Handlungsrelevanz erfolgen. Daher müsste im nächsten Schritt die Erstellung ein Evaluationsplan beauftragt werden. Dieser soll unter anderem die genauen Evaluationsfragestellungen, die Untersuchungsdesigns, die vorgesehenen Datenquellen sowie ein Wirkmodell des Weiterentwicklungskonzepts beinhalten. Darstellung und Differenzierung der verschiedenen Mechanismen des Weiterentwicklungskonzepts helfen dabei, dessen Wirkungen und dessen Komponenten mit den zugrunde liegenden theoretischen Zusammenhängen zu verknüpfen, um ein möglichst umfassendes Bild von der Wirkweise des Weiterentwicklungskonzepts zu erhalten. Das Wirkmodell dient zudem als Unterstützung dabei, die relevanten Fragestellungen der Evaluation zu entwickeln (W. K. Kellogg Foundation 2004). Die Entwicklung eines detaillierten Wirkmodells kann somit erst mit der Erstellung eines Evaluationsplans erfolgen.

10.1 Evaluationsdesign

Je nach Schwerpunkt einer Evaluation liegt das Hauptaugenmerk auf einer formativen oder summativen Evaluation. Für das vorliegende Konzept sind beide Arten von Nutzen. Die genaue Vorgehensweise sowie eine umfassende Festlegung der Schwerpunkte der Evaluation sollten vom Auftraggeber mit dem Augenmerk auf handlungsrelevante Ergebnisse festgelegt werden. Im Folgenden werden beide empfohlenen Evaluationstypen vorgestellt.

Eine prozessbegleitende, formative Evaluation des Weiterentwicklungskonzepts ist zu empfehlen als Maßnahme, um bereits vorgeschlagene Änderungen des Strukturierten Dialogs während dessen Implementierung zu begleiten und gegebenenfalls zu optimieren.

Mögliche Ziele einer formativen Evaluation sind

- Umsetzungsgrad und -hindernisse der einzelnen Maßnahmen zu untersuchen sowie
- die Nützlichkeit (Verständlichkeit, Praktikabilität etc.) der vorgeschlagenen Instrumente aus Sicht der beteiligten Stellen (Landesarbeitsgemeinschaften und Leistungserbringer) zu erfassen.

Bei der Definition von Schwerpunkten der formativen Evaluation ist die Handlungsrelevanz von entscheidender Bedeutung. So kann ein durch eine formative Evaluation festgestelltes Optimierungspotenzial zu einer zusätzlichen Verbesserung der weiterentwickelten Vorgehensweise führen. Beispielsweise könnte eine prozessbegleitende Befragung der Leistungserbringer die Verständlichkeit und Nützlichkeit der entwickelten Empfehlungen zu Form und Inhalt der Einholung von Stellungnahmen erfragen und gegebenenfalls zu Verbesserungsvorschlägen führen.

Eine summative Evaluation nach Umsetzung des vorliegenden Konzepts eröffnet zusätzlich zur Untersuchung von Veränderungen der Prozessqualität (Veit et al. 2013) auch die Untersuchung der Zielerreichung des vorliegenden Konzepts. Einerseits kann in einer summativen Evaluation der Umsetzungsgrad nach Abschluss der Weiterentwicklung gemessen werden. Andererseits kann die Zielerreichung der Weiterentwicklung gemessen werden. Nur bei gleichzeitiger Erreichung eines hohen Umsetzungsgrads und der Erreichung der Ziele der Weiterentwicklung des Strukturierten Dialogs kann von einer erfolgreichen Weiterentwicklung ausgegangen werden. Methodisch wäre sowohl eine retrospektive Beobachtungsstudie als auch eine randomisiert kontrollierte Vergleichsstudie denkbar. Letztere ist trotz des Vorteils der Unabhängigkeit von zeitabhängigen Störgrößen nicht zu empfehlen, da mit der Randomisierung ein deutlich erhöhter Aufwand bei den LAG entstehen würde. Aus diesem Grund wird eine retrospektive Beobachtungsstudie empfohlen. Der Effekt der Weiterentwicklung wird dabei in einem Vorher-Nachher-Vergleich ermittelt. Dabei wird der Zustand nach der Weiterentwicklung sowie der Zustand vor jeglicher Veränderung erfasst und die Differenz berechnet. Zeitlich unabhängige individuelle Störgrößen werden so eliminiert. Somit kann sich die Evaluation auf die zeitliche Veränderung bei Einführung der Weiterentwicklung konzentrieren. Damit kann, unter der Annahme der Abwesenheit von zeitabhängigen Störgrößen, von einem kausalen Zusammenhang ausgegangen werden.

Im Folgenden werden mögliche Schwerpunkte einer summativen Evaluation des ersten Teils des Weiterentwicklungskonzepts dargelegt. Hierfür bieten sich in erster Linie die in Abschnitt 2.5 beschriebenen und im dazugehörigen Anhang 4 vertieften Fragestellungen zur Heterogenität in der Einstufung von Leistungserbringern durch die für die Durchführung des Strukturierten Dialogs beauftragten Stellen an. Daraus ergeben sich folgende mögliche Evaluationsfragestellungen:

- **Fragestellung 1**: Wie groß ist die bewertungsstellenabhängige Heterogenität bezüglich der Aufnahme eines Stellungnahmeverfahrens und hat sich diese durch die Weiterentwicklung reduziert?
- **Fragestellung 2**: Wie groß ist die bewertungsstellenabhängige Heterogenität bezüglich der Bewertung als qualitativ auffällig nach Aufnahme eines Stellungnahmeverfahrens und hat sich diese durch die Weiterentwicklung reduziert?

Konkret empfiehlt sich eine Auswertung basierend auf dem Regressionsmodell in Formel 1 des zweiten Abschnitts in Kapitel 4 des Anhangs. Von primären Interesse für die Evaluation ist dabei die Entwicklung der jeweiligen Heterogenität im Vergleich zum Status vor der Weiterentwicklung. Dabei ist zu empfehlen, das Regressionsmodell auf alle Qualitätsindikatoren eines Leistungsbereichs zu erweitern, da damit die statistische Teststärke (Power) steigt. Die im vierten

Kapitel des Anhangs beschriebenen Methodik lässt sich in dieser Hinsicht erweitern. Eine Messung aller Leistungsbereiche sollte das Problem multipler Messungen berücksichtigen, ist aber zu empfehlen, um die Unsicherheit der Messung des zeitlichen Trends der Heterogenität weiter zu reduzieren. Die Operationalisierung dieser Fragestellungen mithilfe von Evaluationskennziffern würde auf den Ausführungen des vierten Kapitels des Anhangs basieren. Dies sollte in einem zu erstellenden Evaluationsplan dargestellt werden.

Darüber hinaus sollten weitere handlungsrelevante Schwerpunkte wie etwa Fragestellungen zum Ziel der Transparenz und Nachvollziehbarkeit der Entscheidungsfindung definiert und anhand eines Wirkmodells im Rahmen des Evaluationsplans ausgearbeitet werden.

10.2 Datenquellen und Zeitschiene

Naheliegende routinemäßig verfügbare Datenquellen sind die Indikatorergebnisse der Qualitätssicherung und die Bewertungsergebnisse im Modul „Qualitätsbewertung“. Darüber hinaus ist die Befragung der LAG sowie der Leistungserbringer, inklusive eines Zeitpunktes vor der Weiterentwicklung, zu empfehlen.

Eine formative Evaluation sollte prozessbegleitend mit der Implementation der Empfehlungen stattfinden, eine summative Evaluation nach Abschluss der Implementation. Einerseits gilt es, den Zustand nach dem Einpendeln der Implementation, d. h., bei hinreichendem Umsetzungsgrad zu messen. Andererseits sollte der zeitliche Abstand einer Vorher-Nachher-Messung bei einer retrospektiven Beobachtungsstudie und das damit einhergehende Risiko einer Verzerrung durch zeitveränderliche Störgrößen minimiert werden. Unter Berücksichtigung des zu messenden Umsetzungsgrades und des Umstands, dass dadurch die Effekte der Weiterentwicklung möglicherweise noch nicht in vollem Umfang zum Tragen gekommen sind, wird die Messung im ersten Kalenderjahr nach Einführung der Weiterentwicklung empfohlen.

10.3 Zusammenfassung der Empfehlungen zur Evaluation des vorliegenden Konzepts

Es wird empfohlen, sowohl eine formative als auch eine summative Evaluation des vorliegenden Konzepts durchzuführen. Die formative Evaluation soll die Weiterentwicklung unterstützen, indem prozessbegleitend der Umsetzungsgrad der Empfehlungen und die Nützlichkeit aus Sicht der Beteiligten gemessen werden sollen. Die summative Evaluation im Rahmen einer retrospektiven Kohortenstudie ist geeignet, den Umsetzungsgrad und die Zielerreichung im Vergleich zu einer Nullpunktmessung zu erfassen. Empfohlene Evaluationsschwerpunkte sind unter anderem die Untersuchung der Veränderung der Heterogenität in der Einstufung von Leistungserbringern durch die LAG. Es wird empfohlen, im Rahmen der Erstellung eines Evaluationsplans, ein detailliertes Wirkmodell zu erarbeiten, um alle Aspekte der Wirkweise des vorliegenden Weiterentwicklungskonzepts auch in dessen Umfeld in vollständiger Weise untersuchen zu können.

10.4 Evaluation der quantitativen Qualitätsbewertung

In Abschnitt 5.3.1 wurden drei alternative Operationalisierungen des hinreichenden Hinweises für ein Qualitätsdefizit vorgestellt: die rechnerische Auffälligkeitseinstufungsmethode, die statistisch signifikante Einstufungsmethode und die statistisch relevante Einstufungsmethode. In Abschnitt 5.6 wurde empfohlen, die Einstufung nach der statistisch signifikanten Einstufungsmethode vorzunehmen. Des Weiteren wurden in Abschnitt 5.4 Erweiterungen dieser Klassifikationsmethode im Rahmen einer 2-Jahres-Einstufung vorgeschlagen. Im Folgenden werden die empfohlene 1-Jahres-Klassifikation nach der statistisch signifikanten Auffälligkeit sowie die 2-Jahres-Version dieser Einstufungsmethode exemplarisch anhand der Indikatorergebnisse der QS-Verfahren *Hüftendoprothesenversorgung* (HEP) und *Nierentransplantation* (NTX) auf Basis der Ergebnisse des Strukturierten Dialogs für das Erfassungsjahr 2017 verglichen. Ziel dieses Kapitels ist es dabei, die Auswirkungen der gewählten Einstufungsmethode analog zu Abschnitt 2.6.1 abzuschätzen, um zu illustrieren, welche Aufwandsreduktion (im Sinne einer Reduktion der quantitativen Auffälligkeiten) erreicht werden würde. Darüber hinaus werden aufwandsorientierte Vorschläge für den in der Klassifikationsmethode zu wählenden „Tuning-Parameter“ in Form des Signifikanzniveaus erarbeitet. Für die Illustration wurden dabei ein QS-Verfahren mit einer großen Anzahl an Leistungserbringern und behandelten Fällen (HEP) sowie ein Verfahren mit wenigen Leistungserbringern und behandelten Fällen (NTX) ausgewählt.

Der Vergleich dient zum einen dazu, eine Abschätzung vorzunehmen, wie die Anzahl an quantitativen Auffälligkeiten von dem Tuning-Parameter der Einstufungsmethode abhängt. Weiterhin wird untersucht, wie viele der rechnerischen Auffälligkeiten aus dem Erfassungsjahr 2017, die anschließend im Strukturierten Dialog als „qualitativ auffällig“ bewertet wurden, auch anhand der hier untersuchten Methode als statistisch signifikant auffällig eingestuft worden wären. Die Klassifikationsmethode sowie die 2-Jahres-Einstufung werden im Folgenden hinsichtlich der folgenden vier Aspekte betrachtet:

- Die über den vom Referenzwert hinaus tolerierte Anzahl an Fällen mit unerwünschten Ereignissen, die nicht im Strukturierten Dialog analysiert wurden bzw. worden wären, da das Leistungserbringerergebnis nicht als quantitativ auffällig eingestuft wird
- Die Anzahl der quantitativen Auffälligkeiten
- Die Anzahl an qualitativen Auffälligkeiten, die durch die quantitative Einstufungsmethodik vorhergesagt wird (unter der Annahme, dass der darauffolgende Bewertungsprozess dem bisherigen SD-Prozess entspricht)
- Positive-Predictive-Value (PPV) als Maß für die Treffsicherheit des Einstufungsalgorithmus gegeben der bisherigen SD-Ergebnissen

Die Auswertungen werden dabei jeweils differenziert danach, ob es sich um Indikatoren mit festem oder verteilungsabhängigem Referenzbereich handelt. Für verteilungsabhängige Referenzbereiche wird für die statistisch signifikante Auffälligkeitseinstufung der Referenzwert so gewählt, dass immer $q \cdot 100\,\%$ aller Standorte quantitativ auffällig werden, wobei q dem gewählten Anteil an quantativen Auffälligkeiten für den Indikator entspricht – vgl. Abschnitt

5.3.4. Da die Anzahl an quantitativen Auffälligkeiten unabhängig vom Tuning-Parameter konstant ist, ist eine Aufwandsreduktion im Sinne einer Reduktion der quantitativen Auffälligkeiten hier – anders als bei Indikatoren mit festem Referenzwert – nicht möglich.

10.4.1 Limitationen der Evaluation

Bei der folgenden Evaluation ist grundsätzlich zu beachten, dass die qualitative Auffälligkeitseinstufung im Rahmen des Strukturierten Dialogs nicht als Goldstandard für die Bewertung der Qualität eines Leistungserbringers betrachtet werden sollte, da das Ergebnis der qualitativen Bewertung nicht nur von der Qualität der Leistungserbringung sondern auch von anderen Faktoren abhängt, sodass die Bewertung auch eine deutliche bewertungsstellenabhängige Heterogenität ausweißt (vgl. auch Abschnitt 2.5). Schon allein diese Heterogenität im Vorgehen zwischen den den Strukturierten Dialog durchführenden Stellen zeigt, dass es unterschiedliche Vorgehensweisen und Maßstäbe für die qualitative Bewertung zu geben scheint, statt eines einheitlichen Goldstandards.

Außerdem erscheint es plausibel, dass im bisherigen Vorgehen teilweise eine Qualitätsbewertung vorgenommen werden wird, die über das Qualitätsmerkmal des jeweiligen Indikators hinausgeht. So ist es beispielsweise denkbar, dass die durchführenden Stellen auch Ergebnisse des Leistungserbringers in anderen Qualitätsindikatoren für die Bewertung heranziehen, sowie Ergebnisse aus den Vorjahren. Sind beispielsweise im Rahmen eines Strukturierten Dialogs Zielvereinbarungen oder qualitätsverbessernde Maßnahmen mit einem Leistungserbringer vereinbart worden, ist es auch denkbar, dass bei der qualitativen Bewertung im anschließenden Jahr berücksichtigt wird, dass diese Maßnahmen in der Regel eine zeitversetzte Wirkung haben und deshalb auf ein Stellungnahmeverfahren im direkt anschließenden Erfassungsjahr verzichtet wird. Sollte also im bisherigen Vorgehen teilweise eine Qualitätsbewertung vorgenommen worden sein, die über das Qualitätsmerkmal des jeweiligen Indikators hinausgeht, führt dies in Einzelfällen zu Einstufungsergebnissen, die größtenteils unabhängig von dem numerischen Indikatorergebnis sind. Bei einem grundsätzlich validen Indikator und einem Stellungnahmeverfahren, das sich auf das dem jeweiligen Indikator zugrunde liegende Qualitätsmerkmal bezieht, sollte es über alle Stellungnahmeverfahren hinweg einen statistischen Zusammenhang zwischen numerischem Indikatorergebnis und der Wahrscheinlichkeit für eine qualitativ auffällige Bewertung geben. Zeigt sich ein solcher Zusammenhang nicht, ist das Ergebnis der Qualitätsbewertung unabhängig vom Indikatorergebnis. Die damit einhergehenden Probleme sind in Abschnitt 3.3 beschrieben. Der Vergleich der verschiedenen statistischen Operationalisierungen des hinreichenden Hinweises für ein Qualitätsdefizit mit den bisherigen Ergebnissen des Strukturierten Dialogs ist daher auch nur bedingt aussagekräftig

Darüber hinaus ist zu beachten, dass die Sensitivität und Spezifität der qualitativen Bewertung vor dem Hintergrund begrenzter Ressourcen bei den durchführenden Stellen und den Fachkommissionen vermutlich stark von dem Gesamtaufwand aller Dialoge abhängt. Deshalb könnte beispielsweise eine Reduktion der quantitativen Auffälligkeiten selbst schon zu einer Erhöhung der Sensitivität und Spezifität der qualitativen Einstufung führen.

Aus diesen Gründen sollte die vorliegende Evaluation der vorgeschlagenen Klassifikationsmethode nicht als Vergleich gegen einen Goldstandard angesehen werden und die Wahl der Methode und des Tuning-Parameters nicht auf Grundlage dieses Vergleichs, sondern vor allem auf Basis theoretischer Überlegungen getroffen werden. Die retrospektive Evaluation dient daher vor allem dazu, einen Anhaltspunkt für einen Vergleich zum derzeitigen Vorgehen zu ermöglichen und aufzuzeigen, wie groß die Reduktion der quantitativen Auffälligkeiten bei der Wahl verschiedener Werte für den Tuning-Parameter in etwa gewesen wäre, wenn die statistisch signifikante Auffälligkeitseinstufungs-Methodik im Erfassungsjahr 2017 zur Anwendung gekommen wäre, und die anschließende fachliche Bewertung stattdessen der Vorgehensweise des Strukturierten Dialogs für das Erfassungsjahr 2017 entsprochen hätte. Grundsätzlich bleibt offen, ob und inwieweit sich die Evaluationsergebnisse auf die in diesem Bericht vorgeschlagene Neugestaltung des der quantitativen Einstufung nachgelagerten Stellungnahmeverfahrens übertragen lassen.

Außerdem ist zu beachten, dass die Evaluation anhand zweier exemplarisch ausgewählter QS-verfahren vorgenommen wird und die Ergebnisse nicht zwingend auf andere QS-Verfahren übertragbar sind.

10.4.2 1-Jahres-Einstufung

Betrachtet wird als erstes die Anzahl an Fällen mit unerwünschten Ereignissen (über den Referenzbereich hinaus) in Abhängigkeit vom Tuning-Parameter der statistisch signifikanten Auffälligkeitseinstufung. Diese wird definiert als

$$\mathrm{M}_{\mathrm{stat.sig}}(\alpha) = \sum_{i=1}^{I} \mathrm{E}\left[l_i(\cdot)\right] \cdot \left(1 - \mathrm{I}\left(\mathrm{s}_{\mathrm{stat.sig}}(o_i, J_i, R, \alpha)\right)\right),$$

wobei $\mathrm{I}(\cdot)$ die Indikatorfunktion bezeichnet, die den Wert 1 annimmt, wenn ein Leistungserbringerergebnis statistisch signifikant auffällig wird. $l(\cdot)$ ist wiederum die über die vom Referenzbereich hinaus tolerierte Anzahl an Fällen mit unerwünschten Ereignissen hinausgeht: $(\theta - R)_+ \cdot J$.

$\mathrm{M}_{\mathrm{stat.sig}}$ ist damit die über alle Leistungserbringer erwartete Anzahl an Fällen, die über die vom Referenzbereich hinaus tolerierte Anzahl an Fällen mit unerwünschten Ereignissen hinausgeht und für die kein Stellungnahmeverfahren durchgeführt wird, da das Leistungserbringerergebnis nicht als quantitativ auffällig eingestuft wird. Die Größe M ist somit ein Maß für die Anzahl an Fällen mit potentiellen Mängeln in der Behandlungsqualität, die jedoch nicht im Stellungnahmeverfahren nachverfolgt werden. Im Fall der rechnerischen Auffälligkeitseinstufung werden alle Leistungserbringer mit mehr unerwünschten Ereignissen als vom Referenzbereich toleriert als quantitativ auffällig eingestuft. Da jedoch im aktuellen Strukturierten Dialog nicht immer eine Stellungnahme angefordert wird, wird auch die Anzahl an Fällen mit unerwünschtem Ereignis, die durch den Verzicht auf das Anfordern einer Stellungnahme trotz rechnerischer Auffälligkeit in Kauf genommen wird, betrachtet. Sie ist definiert als

$$\mathrm{M}_{\mathrm{Hinweis}} = \sum_{i=1}^{I} \mathrm{E}\left[l_i(\cdot)\right] \cdot \mathrm{I}(\mathrm{s}_{\mathrm{rech}}(o_i, J_i, R) \wedge \text{keine Stellungnahme}).$$

Abbildung 43 stellt $\mathrm{M}_{\mathrm{stat.sig}}$ und $\mathrm{M}_{\mathrm{Hinweis}}$ in Abhängigkeit des Tuning-Parameters α dar, der das Signifikanzniveau für die statistisch signifikante Auffälligkeit darstellt. Die oberen beiden Grafiken zeigen den Verlauf für Qualitätsindikatoren mit festem Referenzwert, die unteren beiden Grafiken den für Indikatoren mit perzentilbasiertem Referenzwert. $\mathrm{M}_{\mathrm{stat.sig}}$ und $\mathrm{M}_{\mathrm{Hinweis}}$ wurden dabei gepoolt über alle 11 ratenbasierten Qualitätsindikatoren des QS-Verfahrens *Hüftendoprothesenversorgung* bzw. alle 7 ratenbasierten Qualitätsindikatoren des QS-Verfahrens *Nierentransplantation* berechnet.

In den oberen beiden Grafiken für Indikatoren mit festem Referenzwert wird dabei deutlich, dass die Kurve für $\mathrm{M}_{\mathrm{stat.sig}}$ für $\alpha = 0{,}47$ bei 0 beginnt, da in diesem Fall alle rechnerisch auffälligen Ergebnisse auch mit der statistisch signifikanten Einstufungsmethode als quantitativ auffällig eingestuft werden und dadurch alle Standorte mit mehr unerwünschten Ereignissen als durch den Referenzbereich zugelassen in den Strukturierten Dialog kommen würden (unter der Annahme, es würden immer Stellungnahmen angefordert und keine Hinweise versendet werden). Als Vergleichswert wird die erwartete Anzahl $\mathrm{M}_{\mathrm{Hinweis}}$ dargestellt, die im Strukturierten Dialog zum Erfassungsjahr 2017 durch das Versenden von Hinweisen bzw. den Verzicht auf das Anfordern einer Stellungnahme entsteht. Da diese Anzahl unabhängig vom Tuning-Parameter ist, entsteht in der Abbildung 43 für diese Methode eine Gerade.

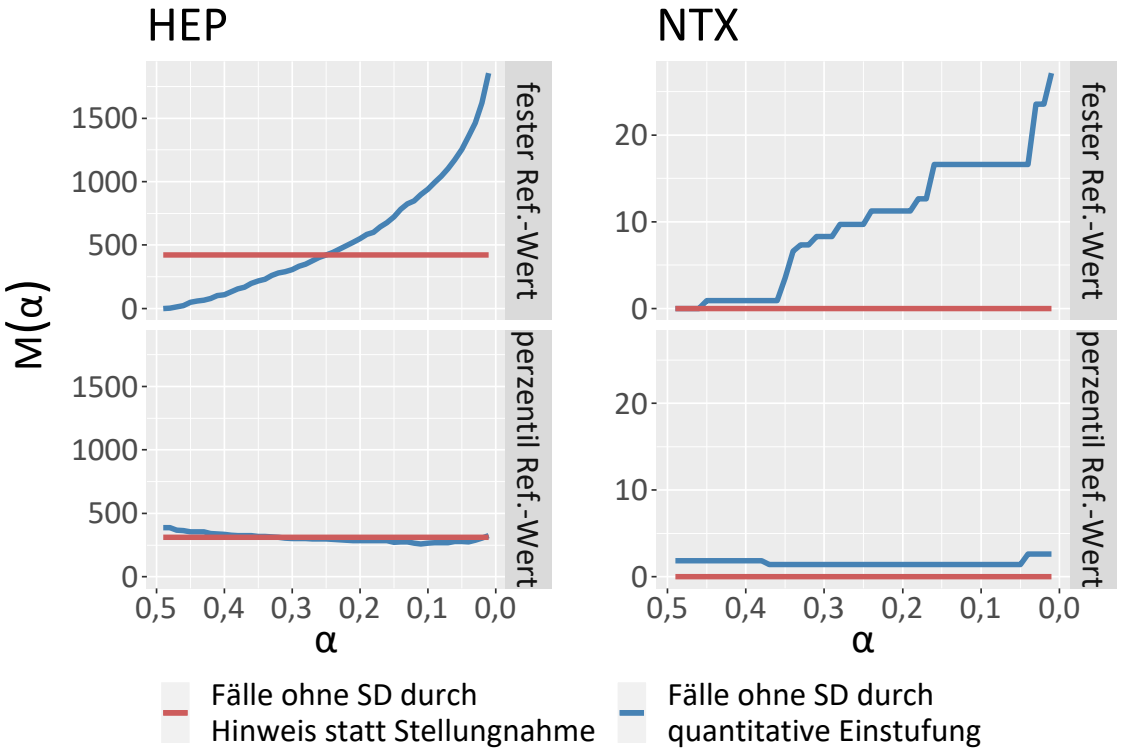

Abbildung 43: Über die vom Referenzwert hinaus tolerierte Anzahl an Fällen mit unerwünschtem Ereignis in Abhängigkeit des Tuning-Parameters ($M(\alpha)$)

Bei festen Referenzwerten zeigt sich: Je mehr statistische Evidenz für eine quantitativ auffälligke Einstufung verlangt wird (d. h. je kleiner α), desto größer ist M. Bei HEP entspricht $\mathrm{M}_{\mathrm{stat.sig}}$ bei etwa $\alpha = 0{,}25$ dem Wert von $\mathrm{M}_{\mathrm{Hinweis}}$, d. h. der Anzahl an Fällen, die in der jetzigen Durchführung des Strukturierten Dialogs durch das Versenden eines Hinweises statt dem Anfordern einer Stellungnahme nach einer rechnerischen Auffälligkeit toleriert wird. Da bei NTX bei jeder rechnerischen Auffälligkeit auch eine Stellungnahme angefordert wird, entspricht $\mathrm{M}_{\mathrm{stat.sig}}$ bei etwa $\alpha = 0{,}47$ dem Wert von $\mathrm{M}_{\mathrm{Hinweis}}$.

Bei perzentilbasierten Referenzwerten, bei denen beispielsweise 5 % der Leistungserbringer quantitativ auffällig werden, hat α nur in Ausnahmefällen einen Einfluss auf M. Da in der rechnerischen Auffälligkeitseinstufung und Berechnung von perzentilbasierten Referenzwerten in der Regel mehr als die nominell erwarteten 5 % der Leistungserbringer auffällig wurden, ist M auch bei großem α zum Teil größer, als die Anzahl, die durch das Versenden von Hinweisen entsteht. Es besteht also ein nicht monotoner Zusammenhang zwischen dem jeweiligen Tuning-Parameter und der Maßzahl M. Dies trifft grundsätzlich auf Qualitätsindikatoren mit perzentilbasiertem Referenzwert zu, da bei diesen die Anzahl an quantitativen Auffälligkeiten festgelegt ist und die Klassifikationsmethoden sich nur darin unterscheiden können, welche (aber nicht wie viele) Leistungserbringerergebnisse quantitativ auffällig werden. Siehe auch Abschnitt 6.4.3 für eine Diskussion der methodischen Limitationen von perzentilbasierten Referenzbereichen.

10.4.2.1 Anzahl an quantitativen Auffälligkeiten

Im Folgenden wird die Anzahl an quantitativen Auffälligkeiten in Abhängigkeit von α dargestellt. In Abbildung 44 wird dabei wieder zwischen ratenbasierten Indikatoren mit festem (in den oberen beiden Grafiken) und perzentilbasiertem Referenzwert (in den unteren beiden Grafiken) unterschieden. Auf der x-Achse ist α dargestellt, auf der y-Achse die absolute Anzahl an quantitativen Auffälligkeiten über alle ratenbasierten Indikatoren. Schwarz eingezeichnet ist die Anzahl an rechnerischen Auffälligkeiten und in Rot die Anzahl an Stellungnahmen im Strukturierten Dialog.

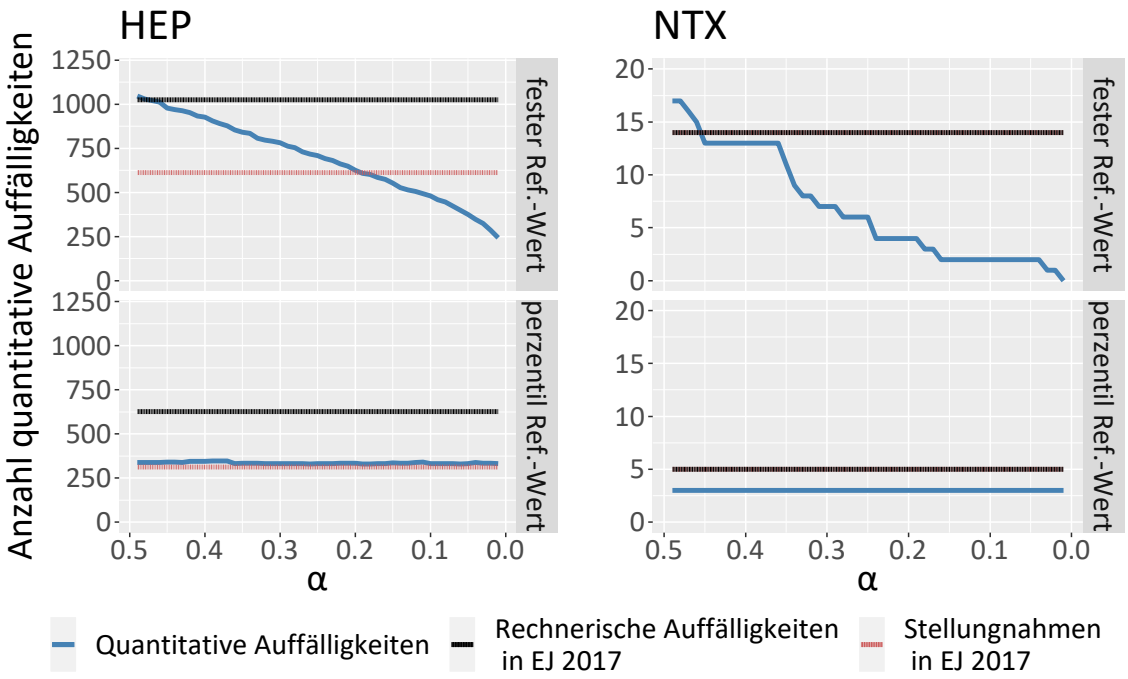

Abbildung 44: Anzahl an quantitativen Auffälligkeiten in Abhängigkeit des Tuning-Parameters

Dabei fällt auf, dass bei perzentilbasierten Referenzwerten unabhängig vom Tuning-Parameter die Anzahl an zu bearbeitenden Stellungnahmen bei HEP nahezu konstant auf dem Wert liegt, der aktuell von den Bewertungsstellen bearbeitet wird, bei NTX wiederum konstant knapp unterhalb dieses Wertes. Dies ist deshalb nicht verwunderlich, da durch die Perzentilreferenzwerte die letztliche Anzahl an quantitativen Auffälligkeiten bereits festgelegt ist.

Bei festen Referenzwerten ist der Einfluss des Tuning-Parameters α auf die Aufwandseinsparung dagegen deutlich erkennbar.

10.4.2.2 Anzahl an qualitativen Auffälligkeiten

Im Folgenden wird verglichen, welcher Anteil an Leistungserbringerergebnissen, die im Anschluss an eine rechnerische Auffälligkeit im Strukturierten Dialog für das Erfassungsjahr 2017 als qualitativ auffällig eingestuft wurden, mit der statistisch signifikanten Klassifikationsmethode als quantitativ auffällig bewertet worden wäre. Abbildung 45 zeigt den Anteil in Abhängigkeit von α, jeweils getrennt nach ratenbasierten Qualitätsindikatoren mit festem Referenzwert (obere Grafiken) und Qualitätsindikatoren mit perzentilbasiertem Referenzwert (untere Grafiken).

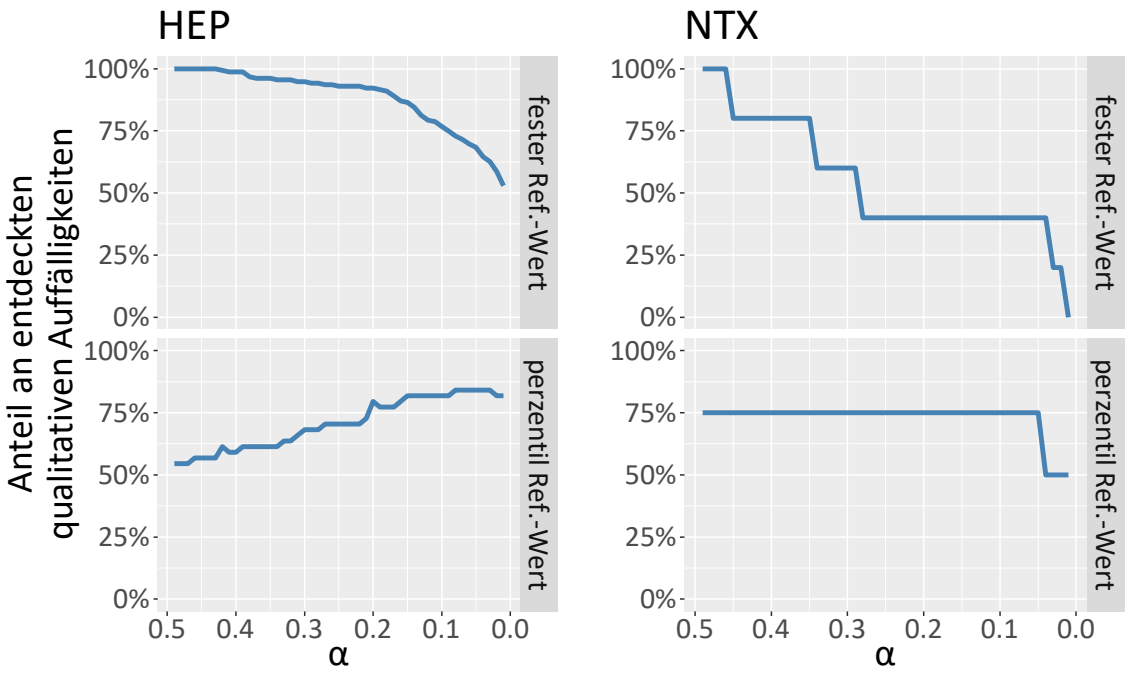

Abbildung 45: Entdeckte qualitative Auffälligkeiten in Abhängigkeit des Tuning-Parameters

Bei HEP würden bei Werten von $\alpha = 0{,}05$ im hypothetischen Vergleich ca. 75 % der qualitativen Auffälligkeiten aus dem SD 2017 auch als quantitativ auffällig eingestuft werden. Bei NTX fällt der Anteil an entdeckten Auffälligkeiten mit kleiner werdendem α bei Indikatoren mit festem Referenzwert deutlich schneller ab und liegt schon bei einem Wert von $\alpha = 0{,}25$ bei unter 50 %. In Anbetracht der in Abschnitt 10.4.1 genannten Limitationen dieser Analyse sind diese Ergebnisse jedoch mit großer Zurückhaltung zu interpretieren. Insbesondere muss bemerkt werden, dass eine strengere Klassifikation, d.h. ein kleineres α, automatisch immer mit einer Reduktion der quantitativen Auffälligkeiten und damit der entdeckten qualitativen Auffälligkeiten gemäß SD 2017 einhergeht.

10.4.2.3 Positive-Predictive-Value

Im Folgenden wird der Positive-Predictive-Value (PPV) für die empfohlene Klassifikationsmethode als Anteil an qualitativen Auffälligkeiten unter allen als quantitativ auffällig eingestuften Leistungserbringern betrachtet.

Für perzentilbasierte Referenzbereiche ist für die Evaluation mit den Ergebnissen des Strukturierten Dialogs aus dem Erfassungsjahr 2017 dabei zu beachten, dass in seltenen Ausnahmefällen Leistungserbringerergebnisse mit der neu vorgeschlagenen Klassifikationsmethode quantitativ auffällig werden, die damals nicht rechnerisch auffällig waren. In diesen Fällen liegt keine Qualitätsbewertung aus dem Strukturierten Dialog zum Erfassungsjahr 2017 vor. Der PPV könnte dementsprechend höher oder niedriger sein; dies wird im Folgenden dadurch illustriert, dass der Bereich zwischen Best- und Worst-Case-Szenario (d. h. alle bzw. keiner der quantitativ auffällig aber nicht rechnerisch auffälligen Standorte wäre auch qualitativ auffällig) farblich dargestellt wird. Es handelt sich dabei nicht um ein Konfidenzband.

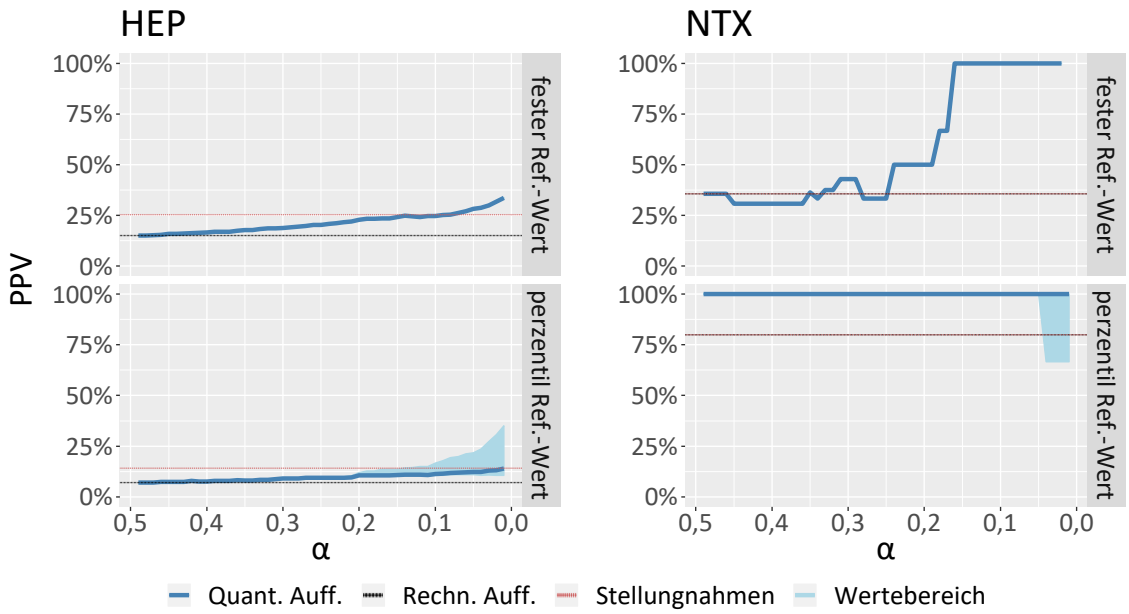

Abbildung 46: PPV in Abhängigkeit des Tuning-Parameters

Bei festen Referenzwerten steigt der PPV mit sinkendem α und ist für HEP niemals geringer als der PPV der rechnerischen Auffälligkeit. Bei NTX ist die Schwankung des PPV über die verschiedenen Werte von α größer und liegt teilweise unter der rechnerischen Auffälligkeit, was unter anderem auch durch die kleine Anzahl an betrachteten Standorten erklärt werden kann. Ab einem $\alpha < 0{,}15$ steigt der PPV im Beispiel NTX jedoch auf 100 %, was bedeutet, dass durch die statistisch signifikante Einstufung ausschließlich Standorte quantitativ auffällig werden, die letztlich auch als qualitativ auffällig bewertet wurden.

Bei HEP entsteht ein Unterschied im PPV abhängig davon, ob alle rechnerischen Auffälligkeiten als Grundlage herangezogen werden, oder lediglich jene, bei denen nachaktuellen Vorgehen gemäß QSKH-RL Stellungnahmen angefordert wurden (vlg. Linien „Rechn. Auff." vs. „Stellungnahmen" in Abbildung 46). Betrachtet man den PPV des aktuellen Vorgehens gemäß QSKH-RL („Stellungnahmen"), so wird deutlich, dass eine ähnliche Steigerung des PPV durch das neue Einstufungsverfahren im Vergleich zur rechnerischen Auffälligkeit erreicht werden kann. Bei der Wahl von $\alpha = 0{,}05$ ist der PPV bei Indikatoren aus HEP mit festem Referenzwert bei über 25 % und damit fast doppelt so groß, wie der PPV der rechnerischen Auffälligkeit, jedoch nur knapp über dem PPV der Stellungnahmen.

10.4.2.4 Abwägung von Aufwandsreduktion und Sensitivität

Der oben grafisch dargestellte Vergleich der Methoden zur aktuellen rechnerischen Auffälligkeit mit den durchgeführten Stellungnahmeverfahren im Erfassungsjahr (EJ) 2017 ist in Tabelle 16 für die statistisch signifikante Auffälligkeitseinstufungsmethode mit ausgewählten Werte von α noch einmal tabellarisch für ratenbasierte HEP-Indikatoren mit festen Referenzbereichen illustriert. Da diese Tabelle nur eine vereinfachende Zusammenfassung der oben dargestellten Grafiken ist, wird hier auf die Darstellung des Leistungsbereichs NTX verzichtet.

Tabelle 16: Stellungnahmeverfahren (STNV), quantitative Auffälligkeiten, entdeckte qualitative Auffälligkeiten und PPV für ausgewählte Werte von α

α	Durchgeführte STNV für das EJ2017	quantitative Auffälligkeiten (stat.sig. Auffälligkeit)	Verhältnis quant. Auffälligkeiten zu STNV (in %)	Anteil an vorhergesagten qualitativen Auffälligkeiten des EJ2017	PPV der quantitativen Auffälligkeit
0,20	613	629	103 %	92 %	23 %
0,10	613	484	79 %	77 %	25 %
0,05	613	378	62 %	68 %	28 %

Wäre beispielsweise die statistisch signifikante Auffälligkeit mit $\alpha = 0{,}2$ als Klassifikationsmethodik verwendet worden, wären im Erfassungsjahr 2017 genau 629 Leistungserbringerergebnisse quantitativ auffällig geworden, was in etwa der Anzahl an durchgeführten Stellungnahmeverfahren entspricht. Bei diesem Signifikanzniveau wären außerdem 92 % der später als qualitativ auffällig bewerteten Ergebnisse auch durch die quantitative Auffälligkeit vorhergesagt worden. Bei $\alpha = 0{,}05$ verringert sich die Anzahl an quantitativen Auffälligkeiten auf etwa 62 % der durchgeführten Stellungnahmeverfahren und es wären noch 68 % der als qualitativ auffällig bewerteten Ergebnisse durch die quantitative Auffälligkeit vorhergesagt geworden. Dabei ist, wie oben beschrieben, grundsätzlich zu beachten, dass die qualitative Auffälligkeit keinen Goldstandard der tatsächlichen Qualitätsbewertung darstellt. Außerdem ist zu berücksichtigen, dass diese Werte in Abhängigkeit des betrachteten QS-Verfahrens stark variieren, wie der oben dargestellte Vergleich der Verfahren HEP und NTX illustriert. Daher kann es unter Umständen sinnvoll sein, den Tuning-Parameter α verfahrens- oder QI-spezifisch zu wählen.

10.4.3 Fazit

Die Evaluation der empfohlenen Einstufungsmethodik anhand der QS-Verfahren HEP und NTX gibt Anhaltspunkte für die zu erwartende Anzahl an quantitativen Auffälligkeiten und vorhergesagter qualitativer Auffälligkeiten in Abhängigkeit des Tuning-Parameters α. Im bisherigen Strukturierten Dialog wurde eine Abwägung zwischen der Anzahl an quantitativen Auffälligkeiten und dem Aufdecken qualitativer Auffälligkeiten zumindest schon teilweise durch das Versenden von Hinweisen in manchen QS-Verfahren, wie z. B. in HEP, getroffen. Obige Analysen zeigen, dass durch eine entsprechende Wahl des Tuning-Parameters die quantitativen Auffälligkeiten auf etwa jene Anzahl an Standorten reduziert werden könnte, von der auch im bisherigen Vorgehen eine Stellungnahme angefordert wurde. Dies würde gegenüber der bisherigen Praxis zu einer Vereinheitlichung über die Bundesländer und Qualitätsindikatoren zu führen. So wäre eine Vereinheitlichung im Vorgehen bei Beibehaltung des bisherigen Volumens an Stellungnahmen erreicht. Aufgrund der Heterogenität im bisherigen Vorgehen würde dies jedoch zu Volumenerhöhung für diejenigen Stellen führen, die bisher mehr Hinweise versendet haben als der Durchschnitt und es käme zu einer Volumenreduktion für diejenigen Stellen, die bisher weniger Hinweise versendet haben. Eine noch größere Volumenreduktion der quantitativen Auffälligkeiten ist darüber hinaus möglich, wenn noch strengere Werte als $\alpha = 0{,}2$ gewählt werden.

Die qualitative Auffälligkeitseinstufung im Rahmen des Strukturierten Dialogs stellt bei dieser Evaluation dabei keinen Goldstandard für die Bewertung der zugrunde liegenden Qualität eines Leistungserbringers dar. Dies ist allein schon an der Heterogenität des Vorgehens der verschiedenen, den Strukturierten Dialog durchführenden Institutionen auf Landesebene erkennbar, welche nahe legt, dass die Maßstäbe für eine qualitative Auffälligkeit stark variieren (vgl. Abschnitt Heterogenitätsanalyse Abschnitt 2.5). Die vorgeschlagene Methodik für die quantitative Auffälligkeitseinstufung kann daher anhand dieser qualitativen Bewertungen nicht gegen einen Goldstandard evaluiert werden und die Wahl des Tuning-Parameters α sollte nicht nur auf dem oben dargestellten Vergleich zur gegenwärtigen Praxis getroffen, sondern auch von theoretischen Überlegungen gestützt werden. Der obige Vergleich gibt jedoch Anhaltspunkte für die Anzahl an quantitativen Auffälligkeiten und vorhergesagter qualitativer Auffälligkeiten der empfohlenen Methode in Abhängigkeit vom verwendeten Tuning-Parameter.

In QS-Verfahren, in denen jede rechnerische Auffälligkeit zu einem Stellungnahmeverfahren geführt hat, führt die empfohlene Einstufungsmethodik immer zu einer Aufwandsreduktion, sofern ein α von kleiner 0,47 gewählt wird, da dieser Wert in etwa der rechnerischen Auffälligkeitseinstufungsmethode entspricht. Die Abwägung zwischen einer Aufwandsreduktion und dem Verlust an entdeckten qualitativen Auffälligkeiten sollte dabei QS-verfahrensspezifisch (und möglicherweise QI-spezifisch) getroffen werden, d. h. der Tuning-Parameter α sollte Verfahrens- und möglicherweise QI-spezifisch wählbar sein.

10.4.4 Mehrjahreseinstufung

Analog zur Klassifikationseinstufung basierend auf einem Erfassungsjahr wird im Folgenden die für die sequentielle Klassifikationsmethodik zweier Erfassungsjahre vorgeschlagene Methodik anhand der Indikatorergebnisse und der Ergebnisse des Strukturierten Dialogs der QS-Verfahren *Hüftendoprothesenversorgung* und *Nierentransplantation* für die Erfassungsjahre 2016 und 2017 illustriert. Wie in Abschnitt 5.4.2 dargelegt, wird ein Leistungserbringerergebnis dabei dann als quantitativ auffällig eingestuft, wenn es entweder im aktuellen Erfassungsjahr statistisch signifikant auffällig ist, oder es im aktuellen Erfassungsjahr rechnerisch auffällig ist und die gepoolten Daten aus dem aktuellen Erfassungsjahr und dem Vorjahr statistisch signifikant auffällig sind[76].

Wie bei der Evaluation der 1-Jahres-Einstufung werden auch hier wieder ratenbasierte Qualitätsindikatoren mit festem Referenzbereich von jenen mit verteilungsabhängigem Referenzbereich getrennt betrachtet und dargestellt. Verteilungsabhängige Referenzbereiche mit Perzentilwert $q \cdot 100\ \%$ werden dabei, wie in Abschnitt 5.4.3 dargelegt, so bestimmt, dass im aktuellen Erfassungsjahr 2017 $q \cdot 100\ \%$ der Leistungserbringerergebnisse statistisch signifikant auffällig werden.

[76] Wobei mit statistisch signifikanter Auffälligkeit die Bayesianische Version $s_{\text{stat.sig.bayes}}$ aus Abschnitt 5.3.1 gemeint ist

10.4.4.1 Anzahl an unerwünschten Ereignissen ohne Stellungnahmeverfahren

Abbildung 47 zeigt die Anzahl an Fällen mit unerwünschten Ereignissen (über den Referenzbereich hinaus), die nicht im Stellungnahmeverfahren diskutiert werden, in Abhängigkeit vom Tuning-Parameter α für die sequentielle Entscheidungsstrategie.

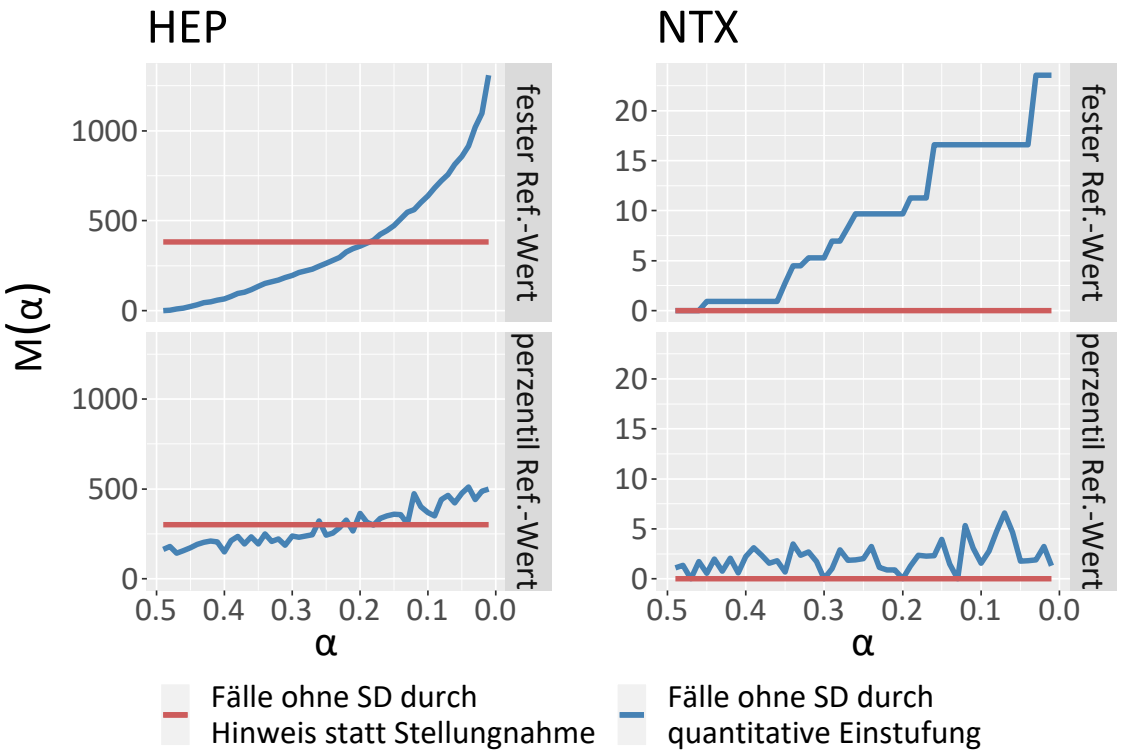

Abbildung 47: Über die vom Referenzwert hinaus tolerierte Anzahl an Fällen mit unerwünschtem Ereignis in Abhängigkeit vom Tuning-Parameter

Betrachtet werden hier nur Standorte, für die auch ein Vorjahresergebnis vorliegt, weshalb die Anzahl an Leistungserbringern ohne Strukturierten Dialog aufgrund der Versendung von Hinweisen statt Stellungnahmeanforderungen etwas niedriger ist als bei der einjährigen Einstufung. Insgesamt ist der Zusammenhang zwischen dem Tuning-Parameter α und der Maßzahl M sehr ähnlich wie bei der einjährigen Einstufung über die statistisch signifikante Auffälligkeit. Für feste Referenzbereiche steigt die Anzahl an überschüssigen Fällen mit unerwünschtem Ereignis, für die kein Stellungnahmeverfahren eingeleitet wird, mit kleiner werdendem α. Bei HEP entspricht bei etwa $\alpha = 0{,}2$ M der Anzahl, die durch das Versenden von Hinweisen derzeit schon in Kauf genommen wird. Dieser Schnittpunkt liegt bei der einjährigen Einstufung bei $\alpha = 0{,}25$, d. h. bei gleichem α führt die sequentielle Einstufung zu einem kleineren M. Dies ist dadurch zu erklären, dass bei der 2-Jahres-Einstufung per Design alle Standortergebnisse auffällig werden, die auch in der einjährigen Einstufung quantitativ auffällig sind, sowie zusätzlich jene Standortergebnisse, deren aktuelles Ergebnis rechnerisch auffällig und deren gepoolte Daten aus den EJ 2017 und 2016 statistisch signifikant auffällig sind. Das heißt bei gleichem α werden bei der sequentiellen Einstufung in der Regel mehr Standortergebnisse auffällig, was zu einem kleineren M führt. Auch bei NTX führt die Klassifikation über die 2-Jahreseinstufung bei gleichem α zu etwas kleinerem M gegenüber der einjährigen Klassifikation.

Bei perzentilbasierten Referenzwerten liegt kein monotoner Zusammenhang zwischen α und M vor und für HEP liegt M wieder durchgängig bei etwa dem Wert, der aktuell durch das Versenden von Hinweisen toleriert wird.

10.4.4.2 Anzahl an quantitativen Auffälligkeiten

Wie oben dargelegt führt die betrachtete 2-Jahres-Klassifikation zu mindestens so vielen quantitativen Auffälligkeiten wie die einjährige Einstufung, was in Abbildung 48 bei Qualitätsindikatoren mit festem Referenzwert erkennbar ist. In Einzelfällen kann es dabei jedoch dadurch zu einer geringeren Anzahl an quantitativen Auffälligkeiten gegenüber der einjährigen Einstufung kommen, dass im Folgenden nur Standorte betrachtet werden, für die für beide betrachteten Erfassungsjahre Daten vorliegen.

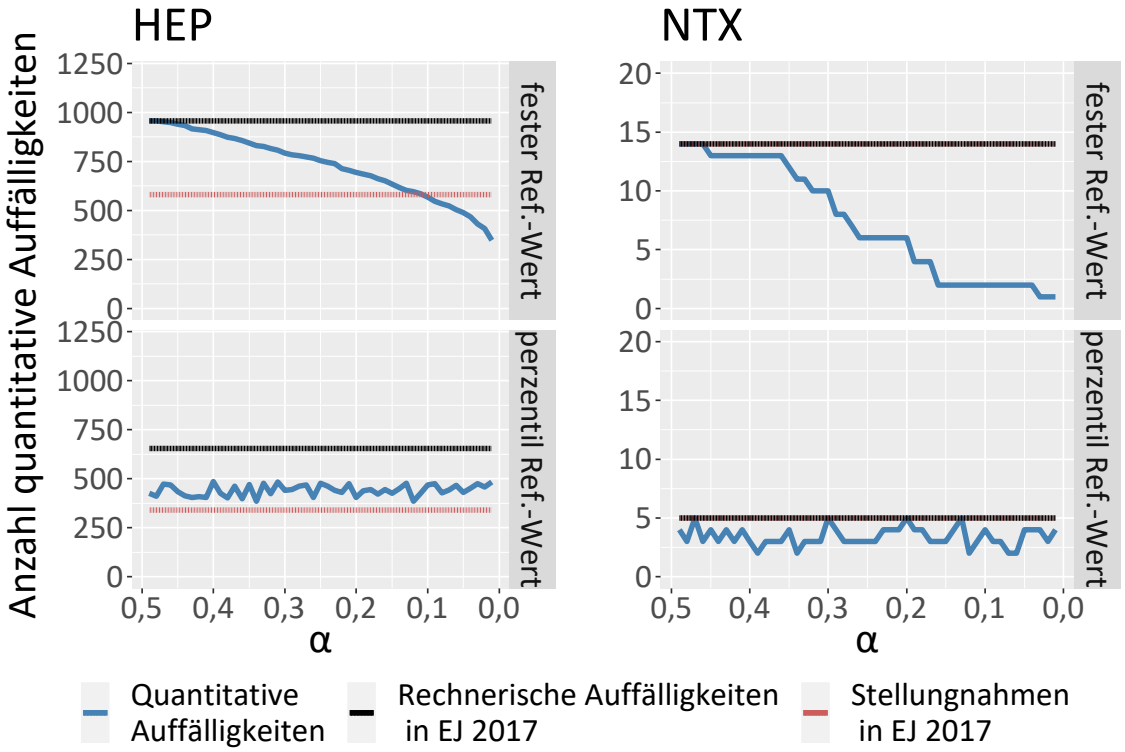

Abbildung 48: Anzahl an quantitativen Auffälligkeiten in Abhängigkeit vom Tuning-Parameter

Für Indikatoren mit festen Referenzwerten nimmt die Anzahl an Stellungnahmeverfahren mit abnehmendem α ab und bei HEP wird bei etwa $\alpha = 0{,}1$ der Wert erreicht, der durch den Versand von Hinweisen erreicht wird. Dies bedeutet, dass die sequentielle Einstufung bei gleichem α zu mehr Stellungnahmeverfahren führt, als die einjährige Klassifikation, aus den oben erläuterten Gründen. Dies gilt auch für NTX, wobei hier die Anzahl an durchgeführten Stellungnahmen wiederum identisch ist zur Anzahl rechnerischer Auffälligkeiten.

Bei Indikatoren mit perzentilbasiertem Referenzwert liegt wieder kein monotoner Zusammenhang zwischen der Anzahl an Stellungnahmeverfahren und dem Parameter α vor. Die Anzahl an quantitativen Auffälligkeiten liegt aber wieder durchgängig über jener der einjährigen Auffälligkeit, da die sequentielle Einstufung per Design zu gleich vielen oder mehr quantitativen Auffälligkeiten führt.

10.4.4.3 Anzahl qualitativer Auffälligkeiten

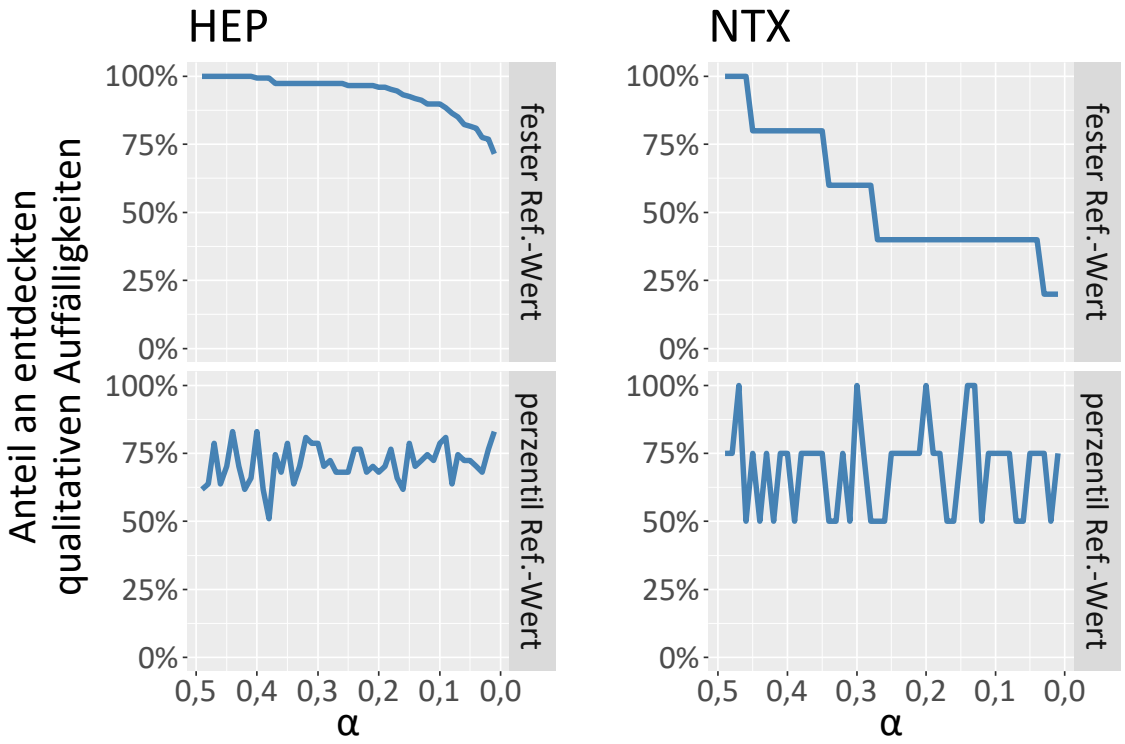

Abbildung 49: Entdeckte qualitative Auffälligkeiten in Abhängigkeit vom Tuning-Parameter

Auch der Zusammenhang zwischen α und dem Anteil an entdeckten qualitativen Auffälligkeiten ist ähnlich wie bei der einjährigen Einstufung über die statistisch signifikante Auffälligkeit. Der Anteil ist bei der sequentiellen Einstufung jedoch durchgängig etwas größer, was wiederum dadurch zu erklären ist, dass die Anzahl an quantitativen Auffäligkeiten bei gleichem α bei der sequentiellen Einstufung größer ist.

10.4.4.4 Positive-Predictive-Value

In Abbildung 50 dargestellt ist der PPV für die sequentielle Einstufung:

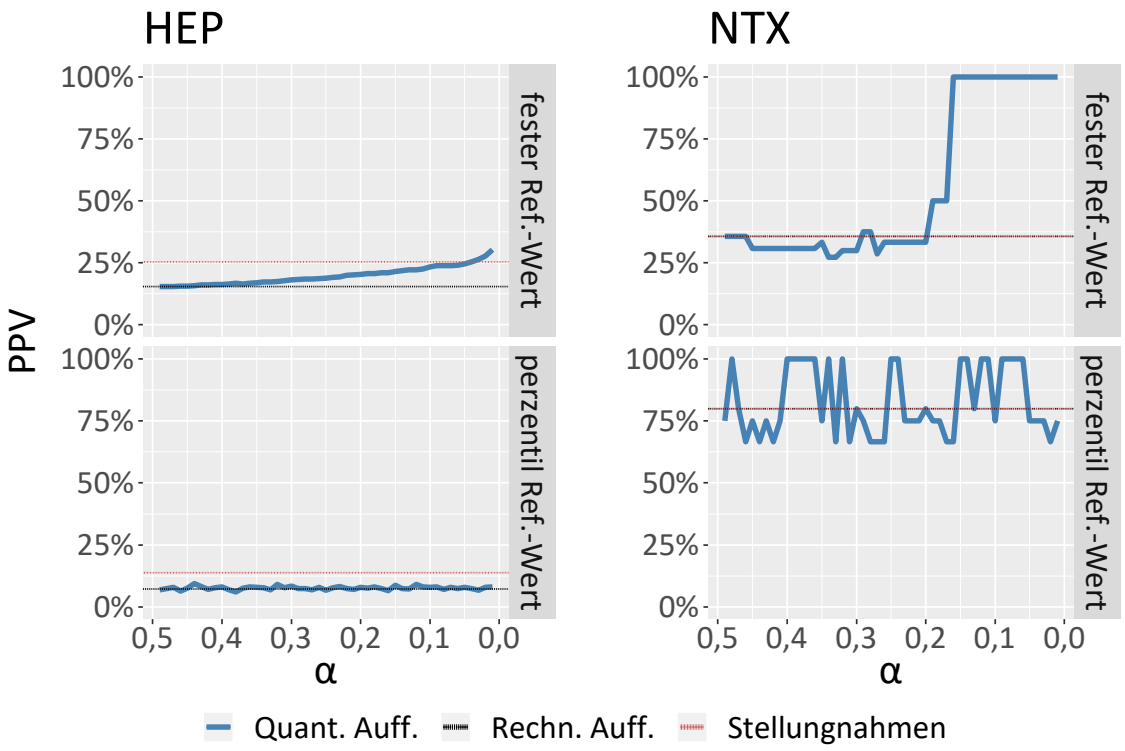

Abbildung 50: PPV in Abhängigkeit vom Tuning-Parameter

Abbildung 50 zeigt, dass der PPV bei HEP durchgängig etwas niedriger ist, als bei der einjährigen Einstufung. Dies bedeutet, dass die zusätzlichen quantitativen Auffälligkeiten durch die gepoolten Daten aus den Erfassungsjahren 2017 und 2016 einen geringeren PPV haben, als die quantitativen Auffälligkeiten aus dem Erfassungsjahr 2017 allein, wenn PPV in der Metrik der SD-Ergebnisse gemessen wird. Dies erscheint insofern plausibel, als dass schwerwiegende Qualitätsmängel im Erfassungsjahr 2017, welches im Strukturierten Dialog als qualitativ auffällig bewertet wurde, schon über die einjährige Einstufung des Erfassungsjahres 2017 quantitativ auffällig werden sollten. Alle *zusätzlichen* quantitativen Auffälligkeiten aus der Zweijahresbetrachtung sind daher insbesondere nicht statistisch signifikant auffällig alleinig auf Basis des Erfassungsjahres 2017 und wurden daher im SD zum Erfassungsjahr 2017 auch seltener als qualitativ auffällig bewertet. Insbesondere wurden gerade diese zusätzlichen quantitativen Auffälligkeiten eventuell häufig nur mit Hinweisen bearbeitet, da einige den Strukturierten Dialog durchführenden Stellen bereits ein Verfahren auf Basis eines statistischen Signifikanztests auf Basis eines Erfassungsjahres für die Durchführung von Stellungnahmen gegenüber dem Versenden von Hinweisen verwenden. Dies trägt dazu bei, dass die Einstufungen aus der Einjahres-Betrachtung einen höheren PPV aufweisen, als die Einstufungen basierend auf zwei Erfassungsjahren. Bei NTX ist dagegen kein deutlicher Unterschied im PPV zwischen der einjährigen und zweijährigen Einstufung zu erkennen.

10.4.4.5 Abwägung von Aufwandsreduktion und Sensitivität

Wie für die 1-Jahres-Einstufung sind in Tabelle 17 die Anzahl an quantitativen Auffälligkeiten und der Anteil an vorhergesagten qualitativen Auffälligkeiten für den Leistungsbereich HEP für ausgewählte Werte von α tabellarisch dargestellt:

α	Durchgeführte Stellungnahme-verfahren	Quantitative Auffälligkeiten (2-Jahres stat. sig. Auffälligkeit)	Verhältnis quant. Auffäligkeiten zu STNV	Anteil an vorhergesagten qualitativen Auffälligkeiten	PPV der quantitativen Auffälligkeit
0,2	580	695	120 %	96 %	20 %
0,1	580	567	98 %	90 %	23 %
0,05	580	488	84 %	82 %	25 %

Tabelle 17: Quantitative Auffälligkeiten, entdeckte qualitative Auffälligkeiten und PPV für ausgewählte Werte von α

Im Vergleich zur 1-Jahres-Klassifikation auf Basis der statistisch signifikanten Auffälligkeit wären bei der 2-Jahres-Einstufung und $\alpha = 0{,}2$ genau 695 Leistungserbringerergebnisse quantitativ auffällig, im Vergleich zu 629 bei der 1-Jahres-Einstufung. Die Zahl der Stellungnahmeverfahren wird dabei bei etwa $\alpha = 0{,}1$ erreicht, wobei 90 % der qualitativ auffälligen Ergebnisse abgedeckt worden wären.

10.4.5 Fazit

Bei der 1-Jahres-Einstufung wurde deutlich, dass eine Aufwandsreduktion im Sinne einer Reduktion der quantitativen Auffälligkeiten gegenüber der aktuellen Anzahl an Stellungnahmeverfahren beim QS-Verfahren HEP erst ab etwa $\alpha = 0{,}2$ erreicht wird. Dies ist dadurch begründet, dass auch bislang nicht für alle rechnerisch auffälligen Standorten eine Stellungnahmen im SD angefordert wurde, sondern stattdessen auch vermehrt lediglich Hinweise verschickt wurden. Ein Reduktion der quantitativen Auffälligkeiten im Vergleich zur Anzahl rechnerisch auffälliger Standorte wird somit dagegen bereits ab einem $\alpha = 0{,}5$ erreicht.

Die Anzahl an quantitativen Auffälligkeiten bei der statistisch signifikanten Einstufungsmethodik über 2 Jahre führt, bei gleicher Wahl des Tuning-Parameters α, zu einer etwas höheren Anzahl an quantitativen Auffälligkeiten als die statistisch signifikante Einstufungsmethodik basierend auf einem Erfassungsjahr. Bei etwa $\alpha = 0{,}1$ wird im Beispiel von HEP die aktuelle Anzahl an durchgeführten Stellungnahmeverfahren erreicht. Der PPV der quantitativen Auffälligkeit ist dabei etwas geringer als jener der 1-Jahres-Einstufung; dies könnte darin begründet sein, dass erhebliche Qualitätsdefizite —die letztlich zu einer qualitativen Auffälligkeit geführt haben — auch bereits zu einer quantitativen Auffälligkeit bei der Betrachtung nur eines Erfassungsjahres führen, da die den Strukturierten Dialog durchführenden Institutionen bereits eine Art der statistischen Auffälligkeit nutzen, um eine Entscheidung zwischen dem Versenden von Hinweisen und dem Führen von Stellungnahmen zu treffen.

Bei NTX führt dagegen jede Wahl von α kleiner als 0,47 zu einer Aufwandsreduktion gegenüber der aktuellen Anzahl an durchgeführten Stellungnahmeverfahren, da dort jede rechnerische Auffälligkeit auch im SD behandelt wird. Auch bei NTX führt die zweijährige Einstufung dabei zu einer etwas höheren Anzahl an quantitativen Auffälligkeiten, als die einjährige Einstufung, bei etwa identischem PPV. Generell sind die hier dargestellten Analyseergebnisse am Beispiel des QS-Verfahrens NTX deutlich wackliger, da bei insgesamt lediglich 14 rechnerischen Auffälligkeiten jede eingesparte Stellungnahme anteilsmäßig einen enormen Effekt hat. Gleichermaßen wiegt auch jede qualitative Auffälligkeit, die somit potentiell nicht entdeckt würde, relativ schwer. Jedoch gilt wie für HEP auch für das verhältnismäßig kleine QS-Verfahren NTX, dass ein kleiner Tuning-Parameter α eine größere Reduktion der ausgelösten Stellungnahmeverfahren und einen tendenziellen Anstieg des PPV in Bezug auf vorliegende Qualitätsmängel bedeutet.

Wie schon bei der Evaluation der 1-Jahres-Klassifikation sollte auch bei der Evaluation der Auswertungsmethodik basierend auf zwei Erfassungsjahren bedacht werden, dass die qualitative Bewertung im Rahmen des Strukturierten Dialogs kein Goldstandard für das tatsächliche Vorliegen von Qualitätsmängeln darstellt und der oben dargestellte Vergleich höchstens eine grobe Abschätzung liefern kann, mit welcher Reduktion an quantitativen Auffälligkeiten gegenüber der gegenwärtigen Praxis eine neue Klassifikationsmethode einhergeht.

11 Zusammenfassungen der Empfehlungen und Konsequenzen

Die vorliegenden Empfehlungen für ein richtlinienübergreifendes Vorgehen bei der Aus- und Bewertung von Indikatorergebnissen sind an den Zielen der Beauftragung ausgerichtet:

- Optimierung der Einheitlichkeit der Vorgehensweise (Objektivität des Verfahrens)
- Optimierung der Transparenz und Nachvollziehbarkeit der Entscheidungsfindung (Objektivität des Verfahrens)
- Optimierung der Effizienz (Standardisierung)

Umgesetzt wurden diese Ziele in Form einer Methodik für die indikatorbasierte Qualitätsmessung und -bewertung, die die weiterführenden Handlungsanschlüsse, wie etwa öffentliche Berichterstattung (*accountability*) und qualitätssteigernde Maßnahmen (*improvement*) berücksichtigt. Das Konzept wurde aus den in Abschnitt 1.2 dargelegten Gründen mit Blick auf die Implementation in der DeQS-RL entwickelt. Durch die modulare Betrachtungsweise lassen sich die Empfehlungen für das Modul „Qualitätsbewertung“ jedoch prinzipiell auch in andere Richtlinien integrieren.

11.1 Anforderungen an und Weiterentwicklung von Qualitätsindikatoren

Die Beauftragung sieht vor dem Hintergrund der gestiegenen Anforderungen an die Prozesse und Ergebnisse der Qualitätssicherung auch eine methodische Weiterentwicklung der bestehenden Qualitätsindikatoren vor. Diese soll laut Beauftragung die Erfahrungen der LAG mittels einer Umfrage berücksichtigen. Eine Überarbeitung der mehr als 200 Qualitätsindikatoren war im Rahmen der Weiterentwicklung des Strukturierten Dialogs jedoch nicht möglich. Eine methodische Überarbeitung ist allerdings auch aus Sicht des IQTIG vor dem Hintergrund der gestiegenen methodischen Anforderungen an die Qualitätsindikatoren erforderlich. Dafür wird nachfolgend ein Vorgehen vorgeschlagen.

Um Qualitätsindikatoren weiterzuentwickeln muss das Ziel der Überarbeitung klar definiert sein. Aus Sicht des IQTIG ist die Maximierung der methodischen Güte (Validität) jedes Qualitätsindikators das Ziel einer solchen Weiterentwicklung. Gemäß den Methodischen Grundlagen V1.1 (IQTIG 2019a) definieren die Eignungskriterien die methodische Güte von Qualitätsindikatoren. Ein Qualitätsindikator ist dann methodisch optimal, wenn er alle Eignungskriterien der Methodischen Grundlagen erfüllt. Je höher die methodische Güte eines Qualitätsindikators ist, desto weniger häufig sollten tendenziell auch besondere Versorgungskonstellationen auftreten, die den Rückschluss vom Indikatorergebnis auf die Versorgungsqualität in Zweifel ziehen. Je seltener solche Konstellationen auftreten, desto geringer fällt auch der Aufwand für das Stellungnahmeverfahren aus, das die Funktion hat, solche Konstellationen zu identifizieren.

Eine methodische Weiterentwicklung der Indikatoren beinhaltet auch die Überprüfung der Referenzbereiche. Da verteilungsbezogene Referenzbereiche keine Aussagen über die Erfüllung definierter Standards erlauben, sondern lediglich vergleichende Aussagen relativ zu den

Ergebnissen aller Leistungserbringer, sind sie für die Qualitätsbewertung nur eingeschränkt nutzbar (vgl. Abschnitt 6.4.3) (IQTIG 2019a, S. 161 f.). Daher sollte im Rahmen der Weiterentwicklung der Indikatoren für alle Indikatoren mit verteilungsbezogenen (perzentilbasierten) Referenzbereichen geprüft werden, ob diese in feste Referenzbereiche überführt werden können.

Für die Weiterentwicklung der Qualitätsindikatoren wird daher ein zweischrittiges Vorgehen empfohlen. Im ersten Schritt werden die Qualitätsindikatoren hinsichtlich der Eignungskriterien geprüft. Für diesen Schritt hat das IQTIG bereits eine Methodik entwickelt, die zur Anwendung im Regelbetrieb kommen soll. Diese Methodik sieht vor, dass das IQTIG unter Einbezug der Bundesfachgruppen eine Einschätzung für jeden Qualitätsindikator hinsichtlich jedes Eignungskriterium vornimmt. Als Ergebnis der Eignungsprüfung jedes Qualitätsindikators steht eine individuelle Einschätzung des Indikators je Eignungskriterium sowie eine zusammenfassende Bewertung mit einer Empfehlung, entweder zur Abschaltung oder zur Überarbeitung.

Dieser Schritt sollte als erster und damit vor der Befragung der LAG erfolgen, da vermutlich schon in diesem Schritt manche Qualitätsindikatoren als nicht mehr geeignet beurteilt werden. So ist beispielsweise für einige Qualitätsindikatoren bekannt, dass das Kriterium „Potenzial zur Verbesserung“ (IQTIG 2019a, S. 140) nicht mehr erfüllt ist. Für diese Indikatoren muss dann keine Rückmeldung von den LAG eingeholt werden. Durch diesen ersten Schritt kann der Aufwand für eine Befragung der 16 LAG zu über 200 Indikatoren reduziert werden. Im zweiten Schritt sollen dann die Erfahrungen der LAG mit den nicht schon im ersten Schritt für die Abschaltung empfohlenen Qualitätsindikatoren eingeholt werden. Diese Erfahrungen sollen dann zusammen mit den Erkenntnissen aus der Eignungsprüfung für die Weiterentwicklung der Indikatoren verwendet werden.

11.2 Gleichsetzung quantitativer Auffälligkeit mit qualitativer Auffälligkeit

Im Rahmen der Beauftragung soll das IQTIG auch Qualitätsindikatoren identifizieren, für die eine rechnerische Auffälligkeit mit einer qualitativen Auffälligkeit gleichgesetzt werden kann (Punkt 3.d der Beauftragung). Dieser Beauftragungspunkt wird so verstanden, dass das IQTIG prüfen soll, unter welchen Umständen ein Stellungnahmeverfahren verzichtbar ist und ob diese Umstände auf bestehende Qualitätsindikatoren zutreffen.

Aus den methodischen Vorüberlegungen und den Gütekriterien für Bewertungsprozesse (vgl. Kapitel 3) ergibt sich die klar umschriebene Funktion des Stellungnahmeverfahrens als nachgeschalteter Prüfschritt zur Sicherstellung der Aussagekraft eines Indikatorergebnisses (Validität je Messung), wenn dieses hinreichende Evidenz für ein Qualitätsdefizit nahelegt (vgl. Kapitel 6). Damit könnte eine Gleichsetzung des quantitativen Indikatorergebnisses mit einer abschließenden Qualitätsbewertung einerseits erfolgen, wenn der Bedarf für diesen Prüfschritt entfiele – also wenn die Aussagekraft eines quantitativen Indikatorergebnisses (im Sinne des Soll-Ist-Vergleichs mit dem Referenzwert ggf. mittels eines statistischen Verfahrens) als ausreichend hoch eingeschätzt würde. Vor dem Hintergrund des Weiterentwicklungsbedarfs der bestehenden Qualitätsindikatoren besteht aus Sicht des IQTIG allerdings weiterhin Bedarf für diesen Prüfschritt (siehe auch Rückmeldungen im Rahmen des Workshops zur Einbindung der Vertreterinnen und Vertreter der LAG und der LQS, Anhang, Kapitel 1).

Andererseits kann eine Entscheidung gegen diesen Prüfschritt auch aus Aufwand-Nutzen-Überlegungen heraus erfolgen, wenn etwa der Aufwand für diesen Prüfschritt nicht durch dessen Nutzen gerechtfertigt erscheint (vgl. Kapitel 5 zu Verlustfunktionen). Um zu einer solchen Einschätzung zu gelangen, müssen Aufwand und Nutzen eines Vorgehens mit diesem Prüfschritt und denen eines hypothetischen Vorgehens ohne diesen Prüfschritt verglichen werden. Der Nutzen dieses Prüfschritts besteht in einer Steigerung der Spezifität des Vorgehens bezogen auf die Wahrscheinlichkeit, „tatsächliche" Qualitätsdefizite zu detektieren im Vergleich zu einem Vorgehen ohne diesen Prüfschritt (vgl. Abschnitt 2.1). Dieser Nutzen kann aus verschiedenen Perspektiven unterschiedlich bewertet werden. Aus Sicht der Leistungserbringer könnte eine höhere Spezifität wünschenswert sein, da die Wahrscheinlichkeit sinkt, irrtümlicherweise unzureichende Qualität attestiert zu bekommen. Aus Sicht von Patientinnen und Patienten könnte die Sensitivität des Verfahrens höher gewichtet werden als die Spezifität, da eher Leistungserbringer mit „tatsächlich" unzureichender Qualität identifiziert würden. Im nächsten Schritt müsste die Frage beantwortet werden, wie groß der Gewinn an Spezifität durch diesen Prüfschritt ausfällt. Dies ist nicht ohne weitere Annahmen möglich. Eine solche weitergehende Analyse geht jedoch über den Rahmen dieser Beauftragung hinaus.

Der Aufwand des Prüfschritts entsteht für verschiedene Beteiligte (Leistungserbringer, LAG bzw. Bundesstelle) durch alle nicht automatisierbaren Schritte des Moduls „Qualitätsbewertung", angefangen von der Einholung der Stellungnahmen (LAG bzw. Bundesstelle), des Verfassens der Stellungnahmen (Leistungserbringer), über deren formale Prüfungen (LAG bzw. Bundesstelle), die inhaltliche Bewertung (Fachexpertinnen und -experten) bis hin zur abschließenden Einstufung des Indikatorergebnisses (siehe auch Kapitel 6). Nicht berücksichtigt ist hier der Aufwand für die Einleitung und Durchführung qualitätsverbessernder Maßnahmen bei den Leistungserbringern, da diese nicht zu dem Modul Qualitätsbewertung zählen. Eine Motivation für die Weiterentwicklung stellt laut Beauftragung auch der hohe Aufwand des bisherigen Vorgehens im zweiten Prüfschritt (dem Stellungnahmeverfahren) sowohl für die Leistungserbringer als auch für die LAG bzw. Bundesstelle dar. Eine teilweise Reduktion dieses Aufwands kann durch methodisch weiterentwickelte Qualitätsindikatoren erreicht werden, da sich dadurch deren Spezifität erhöht und damit der Aufwand für den folgenden Prüfschritt reduziert (siehe auch Abschnitt 11.2). Unter der Annahme, dass die bestehenden Qualitätsindikatoren teilweise eher als Aufgreifkriterien mit eher hoher Sensitivität aber niedriger Spezifität konstruiert wurden, erscheint der Bedarf für einen zweiten, die Spezifität des Vorgehens zur Qualitätsmessung steigernden Prüfschritt jedoch weiterhin gegeben.

Grundsätzlich ergibt sich hier jedoch ein Zielkonflikt zwischen einer maximal fairen Qualitätsmessung einerseits und einem möglichst geringen Aufwand andererseits, vor allem für die Leistungserbringer. Wie in Abschnitt 2.1 dargelegt, erhöht ein Stellungnahmeverfahren, das die Aussagekraft eines auffälligen Indikatorergebnisses überprüft, die Fairness der Qualitätsmessung zugunsten der Leistungserbringer, da es falsch-positive Ergebnisse (siehe Tabelle 1) reduziert und einheitliche a priori definierte Kriterien anlegt. Demgegenüber steht der Aufwand für die Leistungserbringer, für jedes Indikatorergebnis mit dem Hinweis auf ein Qualitätsdefizit eine Stellungnahme zu verfassen. Eine Steigerung der Spezifität und der damit einhergehenden Stei-

gerung der Fairness der Qualitätsmessung (siehe Abschnitt 2.1), wird erkauft durch einen höheren Aufwand im Vergleich zu einem Vorgehen ohne Stellungnahmeverfahren. Diese Aufwandserhöhung durch ein Stellungnahmeverfahren im Vergleich zu einem Vorgehen ohne Stellungnahmeverfahren ist auch im Hinblick auf die vertragsärztlichen und vertragspsychotherapeutischen Leistungserbringer mit dem Gewinn an Fairness durch das Stellungnahmeverfahren abzuwägen. Im Gegensatz zu Leistungserbringern des stationären Sektors verfügen Leistungserbringer des ambulanten Sektors nicht über eigene QM-Abteilungen, die die Erstellung einer Stellungnahme unterstützen können. Eine Steigerung der Spezifität des Vorgehens der Qualitätsbewertung ohne weitere Aufwände bei den Leistungserbringern ist jedoch nicht möglich. Eine Quantifizierung dieser Aufwände ist im Rahmen dieses Projekts allerdings nicht möglich gewesen. Da weder konkrete Nutzen- noch Aufwandsabschätzungen für den Prüfschritt vorgenommen werden konnten, kann hier auch kein Aufwand-Nutzen-Vergleich eines Vorgehens mit und ohne zweiten qualitativen Prüfschritt vorgenommen werden. Anhand dieser Überlegungen wird jedoch deutlich, dass die Entscheidung für oder gegen den zweiten Prüfschritt entscheidend von der Bewertung des Gewinns an Spezifität im Verhältnis zu dem Aufwand abhängig ist. Da sich außerdem sowohl Aufwand als auch Nutzen zwischen den verschiedenen Beteiligten unterscheiden, existieren mindestens zwei verschiedene Perspektiven auf dieses Aufwand-Nutzen-Verhältnis, in Anbetracht der geringeren Personalressourcen pro Leistungserbringer im ambulanten Sektor können vermutlich drei Perspektiven unterschieden werden: die der Patientinnen und Patienten, die der stationären Leistungserbringer und die der ambulanten Leistungserbringer. Vor dem Hintergrund des Weiterentwicklungsbedarfs einiger bestehender Qualitätsindikatoren und dem damit einhergehenden Bedarf an einer (unbekannten) Steigerung der Spezifität des Vorgehens durch den zweiten, qualitatitven Prüfschritt wird empfohlen, das Stellungnahmeverfahren beizubehalten.

11.3 Indizes als Aufgreifkriterien

Die Beauftragung sieht auch vor, dass das IQTIG prüfen soll, ob Indizes als Aufgreifkriterien für die Bewertung ganzer Leistungsbereiche verwendbar sind (Punkt 3.e der Beauftragung). Indizes in diesem Sinn werden verstanden als eine (gewichtete) Aggregation der Ergebnisse mehrere Indikatorergebnisse pro Leistungserbringer und Leistungsbereich. Unter der Annahme, dass bei der Verwendung von Indizes als Aufgreifkriterien für die Bewertung der Leistungserbringung in ganzen Leistungsbereichen keine Betrachtung der einzelnen Indikatorergebnisse mehr nötig ist, könnte dieses Vorgehen zu einer Effizienzsteigerung beitragen, da weniger Ergebnisse bewertet werden müssten. Diese Annahme wird vom IQTIG jedoch nicht geteilt (siehe Abschnitt 4.2).

Wie in Abschnitt 2.2 dargelegt, ist aus Sicht des IQTIG vor dem Hintergrund der gestiegenen Anforderungen an die Qualitätsmessung ein Wandel geboten im Verständnis von Qualitätsindikatoren weg von Aufgreifkriterien hin zu quantitativen Größen für Qualität. Daher sieht das hier vorgeschlagene Konzept vor, dass das Ergebnis jedes einzelnen Qualitätsindikators abschließend bewertet werden soll. Schon heute werden in der gesetzlich verpflichtenden Qualitätssicherung Indikatoren eingesetzt, die selbst ein Index sind (z. B. QI 51901 Qualitätsindex der Frühgebo-

renenversorgung). Bei diesen Indikatoren findet die Qualitätsbewertung jedoch erst auf Indexebene statt – nicht auf Ebene der konstituierenden Kennzahlen oder Variablen. Durch die Entscheidung einen Referenzbereich zu setzen, auf einem Index oder auf einer Kennzahl, fällt per Definition die Entscheidung, auf der jeweiligen Aggregationsebene eine Qualitätsbewertung vorzunehmen. Da in dem hier vorliegenden Konzept die Funktion eines Stellungnahmeverfahrens darin gesehen wird, falsch-positive Qualitätsbewertungen (siehe Abschnitt 2.1) zu reduzieren und damit die Spezifität der Qualitätsbewertung zu erhöhen, folgt aus der Entscheidung einen Referenzbereich zu setzen logisch die Notwendigkeit für ein Stellungnahmeverfahren auf der Aggregationsebene, auf der der Referenzbereich gesetzt wurde.

Dies ist unter anderem auch deshalb notwendig, da die Ergebnisse der einzelnen Qualitätsindikatoren für andere Zwecke weiterverwendet werden. Dieses Vorgehen ist damit auch im Einklang mit dem Vorgehen gemäß plan. QI-RL. Die diesem Punkt der Beauftragung scheinbar zugrunde liegende Annahme, dass eine Bewertung einzelner Indikatorergebnisse verzichtbar sei, wird daher nicht geteilt. Aus diesen Gründen wird eine Verwendung von aggregierten Qualitätsindikatorergebnissen als Aufgreifkriterien nicht empfohlen. Gemäß der Methodischen Grundlagen V1.1 können Indizes jedoch prinzipiell geeignet sein, um die Versorgungsqualität eines Leistungserbringers in einem Leistungsbereich zu einer Messgröße zu verdichten (Profit et al. 2010). Auch für weiterführende Verwendungszwecke, wie etwa eine über einen Qualitätsindikator hinausgehende Qualitätsbewertung, beispielsweise einer Fachabteilung oder wie bei der Entwicklung einer Internetplattform für eine Vergleiche ermöglichende Qualitätsdarstellung, können Indizes geeignet sein.

11.4 Leistungsbereichübergreifende Qualitätsindikatorensets

Des Weiteren beinhaltet die Beauftragung auch die Prüfung von Optionen für leistungsbereichsübergreifende Indikatorensets (Punkt 3.c der Beauftragung). Dieser Punkt der Beauftragung wird so verstanden, dass das IQTIG prüfen soll, ob diagnose- und/oder prozedurunabhängige Qualitätsindikatoren sinnvoll im Rahmen der gesetzlich verpflichtenden QS eingesetzt werden können. Es wird nicht davon ausgegangen, dass solche leistungsbereichübergreifenden Indikatorensets als „Aufgreifkriterien“ verstanden werden.

Das Verständnis von leistungsbereichübergreifenden Indikatoren soll am Beispiel zweier Indikatoren erläutert werden, die jeweils den Anteil von Patientinnen und Patienten mit Komplikationen abbilden, einmal nach Hüftendoprothesenversorgung und einmal nach Knieendoprothesenversorgung. Mittels beider Indikatoren könnten mehr oder weniger für diesen Eingriff spezifische Komplikationen abgebildet werden. Ein leistungsbereichübergreifender Indikator in diesem Beispiel würde den Anteil von Patientinnen und Patienten mit Komplikationen unabhängig von dem Eingriff zusammenfassen. Als Weiterführung dieses Beispiels wäre ein Indikator vorstellbar, der den Anteil von Patientinnen und Patienten mit Komplikationen nach jeglichem Eingriff abbildet. Hier wird deutlich, dass die zugrunde liegende Frage nach diesem Verständnis von leistungsbereichübergreifenden Indikatoren die des Auflösungsniveaus des Indikators ist. Am einen Ende des Kontinuums steht ein für eine bestimmte Prozedur/Diagnose hochspezifi-

scher Indikator (z. B. „Lockerung der Hüftendoprothese“), am anderen Ende steht ein sehr allgemeiner Komplikationsindikator. Eine Effizienzsteigerung im Sinne der Beauftragung würde in diesem Verständnis dadurch entstehen, dass ein Leistungserbringer nur noch in einem statt in zwei Indikatoren auffällig werden könnte und dementsprechend auch nur ein Stellungnahmeverfahren eingeleitet und durchgeführt werden müsste.

Ob ein solches Vorgehen sinnvoll ist, hängt von zwei Faktoren ab: dem Verwendungszweck der Qualitätsmessung und der Annahme, dass das dem Indikator zugrunde liegende Qualitätsmerkmal über die beteiligenden Leistungsbereiche hinweg homogen ist. Für die Ableitung von Maßnahmen zur Qualitätsverbesserung sind spezifischere Indikatoren geeigneter, da sie den Suchraum für die Problemursache eingrenzen. Für Verwendungszwecke im Rahmen von *accountability* können hingegen auch weniger spezifische Indikatoren geeignet sein, beispielsweise um Vergütungszuschläge an allgemein niedrige Komplikationsraten zu knüpfen. Für Patienteninformation zum Zweck der Auswahlentscheidung ist dagegen das Informationsbedürfnis der Patientinnen und Patienten zu berücksichtigen. Denn typischerweise werden Auswahlentscheidungen mit Blick auf einen bestimmten elektiven Eingriff getroffen – der spezifischere Indikator im obigen Beispiel wäre in diesem Fall aussagekräftiger.

Das zweite Kriterium betrifft die Frage, ob angenommen wird, dass das dem Indikator zugrunde liegende Qualitätsmerkmal über die Leistungsbereiche hinweg homogen ist (vgl. auch reflektive vs. formative Konstrukte; Shwartz et al. 2015). Unterscheiden sich in dem obigen Beispiel die Komplikationsraten zwischen den beiden Eingriffen substanziell, können solche Unterschiede in einem allgemeinen Indikator verdeckt werden. Ob eine solche gegenseitige Kompensation eine erwünschte oder eine unerwünschte Eigenschaft eines Indikators ist, hängt unter Umständen von seinem Verwendungszweck ab. Im Rahmen einer Konzeptentwicklung können die Bedingungen für einen sinnvollen Einsatz von leistungsbereichübergreifenden Indikatorensets beschrieben werden.

11.5 Zusammenfassung der zentralen Empfehlungen

Die Anforderungen an die Prozesse und Ergebnisse der gesetzlich verpflichtenden Qualitätssicherung sind in den letzten Jahren deutlich gestiegen (vgl. Abschnitt 2.2). Die Ergebnisse von Qualitätsmessungen sollen beispielsweise für die Planung von Versorgungsstrukturen (plan. QI-RL) oder für die leistungserbringerbezogene Veröffentlichung von Qualitätsergebnissen verwendet werden. Das hier vorgeschlagene Konzept greift die gestiegenen Anforderungen an Qualitätsindikatoren auf und zeigt Wege für eine vor diesem Hintergrund notwendige Weiterentwicklung der Qualitätsbewertung auf.

Mit Hinblick auf das Ziel frühzeitigere und aussagekräftigere Qualitätsergebnisse zu erhalten, wird empfohlen, die Instrumente der Qualitätsbewertung und der Qualitätsförderung konzeptuell und zeitlich voneinander zu trennen. Hierzu wird eine modulare Betrachtungsweise vorgeschlagen, die die Prozesse der Qualitätsbewertung, der Qualitätsförderung und der Bewertung der Dokumentationsqualität voneinander trennt. Durch stärkere Standardisierung soll der bisher teilweise aufwendige und unterschiedlich ausgestaltete Prozess der Qualitätsbewertung ob-

jektiver gestaltet und verschlankt werden. Es wird empfohlen, den Begriff des Stellungnahmeverfahrens nur für den Teil des Vorgehens nach § 17 DeQS-RL zu verwenden, der die Einholung und abschließende Bewertung der Stellungnahmen umfasst. Die darauf aufbauenden Maßnahmen der Qualitätsförderung sollen im Sinne der Trennung von Qualitätsbewertung und Qualitätsförderung nicht als Teil des Stellungnahmeverfahrens verstanden werden. Diese Trennung kann auch dazu beitragen, dass die Qualitätsergebnisse früher verfügbar sind als bisher.

Kriterium für die Einholung von Stellungnahmen

Zur Steigerung der Spezifität und der Effizienz des Verfahrens wird eine statistische Operationalisierung des hinreichenden Hinweises auf ein Qualitätsdefizit auf Basis des quantitativen Indikatorergebnisses empfohlen. Dafür wird in Kapitel 5 ein Rahmenkonzept für die statistische Auswertungsmethodik von Qualitätsindikatoren dargestellt, welches die Festlegung von Herangehensweise, Stichprobenart, Berechnungsart und Bewertungsart erfordert. Bei einer analytischen Herangehensweise sollen stochastische Einflüsse und die damit einhergehende statistische Unsicherheit auf die Ergebnisse berücksichtigt werden. Diese Herangehensweise trifft aus Sicht des IQTIGs für die meisten fallbasierten Qualitätsindikatoren zu. Aufbauend auf dem Konzept wird ein formaler Rahmen beschrieben, innerhalb dessen sich, unter bestimmten Annahmen, die Optimalität der Methode der statistischen Auffälligkeit zeigen lässt. Durch den Einsatz dieser Methode können Verzerrungen bei der Feststellung eines hinreichenden Hinweises auf ein Qualitätsdefizit, wie sie bei Einsatz der einfachen rechnerischen Auffälligkeit bei kleinen Fallzahlen auftreten, vermieden werden. Die bayesianische Version der statistischen Auffälligkeitseinstufung bietet darüber hinaus einen flexibleren Rahmen als derjenige der frequentistischen Auffälligkeitseinstufung, sodass auch komplexere Indikatorarten in einem einheitlichen Rahmen ausgewertet werden können. Auch die Interpretation der Ergebnisse ist deutlich einfacher im bayesianischen Rahmen, z. B. kann von der Wahrscheinlichkeit der Nullhypothese gesprochen werden, d. h. wie wahrscheinlich es ist, nach Beobachtung der Daten, dass der Kompetenzparameter des Leistungserbringers im Referenzbereich liegt. Eine solche bayesianische Auswertungsmethodik wurde z. B. bereits für die Auswertung der Daten von Patientenbefragungen in den Abschlussberichten für die Patientenbefragungen im Verfahren *QS PCI* und im Verfahren *Versorgung von volljährigen Patienten und Patientinnen mit Schizophrenie, schizotypen und wahnhaften Störungen* empfohlen. Für Ratenindikatoren und risikoadjustierte Indikatoren per Poisson-Approximation, wie sie z. Zt. in der plan. QI-RL zum Einsatz kommen, unterscheiden sich die beiden Ansätze nur minimal. Daher empfiehlt das IQTIG auch für die Auswertung der übrigen Indikatoren im Rahmen der DeQS-RL die bayesianische Version der statistischen Auffälligkeit als Kriterium für die Entscheidung, ob ein Leistungserbringer das durch den Referenzbereich definierte Soll in einem Qualitätsindikator erreicht hat.

Durch die Berücksichtigung von Unsicherheit bei der statistischen Entscheidung, ob eine Stellungnahme eingeholt werden soll, wird durch das Signifikanzniveau eine zusätzliche Stellschraube eingeführt, ab wann von hinreichender statistischen Evidenz für ein mögliches Qualitätsdefizit ausgegangen werden kann. Die andere wichtige Stellschraube ist der Referenzwert für den Qualitätsindikator. Bei der in Kapitel 5 vorgeschlagenen statistischen Auffälligkeitseinstufungsmethode wird im bayesianischen Rahmen die Wahrscheinlichkeit berechnet, dass der

latente Kompetenzparameter des Leistungserbringers im Referenzbereich liegt. Das Signifikanzniveau der Auffälligkeitseinstufungsmethode legt dann fest, wie klein diese Wahrscheinlichkeit sein darf, bevor eine Entscheidung für die Alternativhypothese, d. h. der Kompetenzparameter liegt außerhalb des Referenzbereiches, getroffen wird. Die Frage ist, welcher Grad an Evidenz bzw. welches Signifikanzniveau in der Praxis für diese Entscheidung gewählt werden soll. Eine Diskussion dazu findet sich in Abschnitt 5.6 und ist eng mit der Methodik zur Festlegung der Referenzwerte bei den Qualitätsindikatoren verknüpft. Eine sehr pragmatische Vorgehensweise wäre z. B. übergangsweise das Signifikanzniveau so zu wählen, dass in etwa der gleiche Aufwand durch Stellungnahmen resultiert wie im bisherigen Rahmen des Strukturierten Dialogs.

Im Rahmen der DeQS-RL werden alle Auswertungen durch die Bundesauswertungsstelle vorgenommen. Der Aufwand für die Umsetzung der Empfehlungen zur statistischen Operationalisierung ergibt sich somit hauptsächlich für das IQTIG. Da durch die statistische Operationalisierung eindeutig vorgegeben sein soll, für welche Indikatorergebnisse eine Stellungnahme einzuholen ist, bestünde damit kein Ermessensspielraum mehr, wodurch sich ein Gewinn an Objektivität ergibt. Außerdem ergibt sich ein Effizienzgewinn im Vorgehen, da eine zu häufige Einholung von Stellungnahmen bei kleinen Fallzahlen vermieden wird und der Entscheidungsprozess, für welches Indikatorergebnis eine Stellungnahme einzuholen ist, praktisch automatisiert wird.

Durch die statistische Operationalisierung des hinreichenden Hinweises auf ein Qualitätsdefizit ist auch das Versenden von Hinweisen nicht mehr nötig. Bisher wurden Hinweise vermutlich versendet, um bei eher schwachen Hinweisen für Qualitätsdefizite (z. B. 1-Fall-Regel gemäß § 10 QSKH-RL) aufgrund der Auslösung durch die einfache rechnerische Auffälligkeit keine Stellungnahem anzufordern. Da die Leistungserbringer schon durch die Rückmeldeberichte auf ihre Indikatorergebnisse aufmerksam gemacht werden und erwartet wird, dass das interne Qualitätsmanagement diese Ergebnisse von sich aus heranzieht und gegebenenfalls tätig wird, ist das Versenden von zusätzlichen Hinweisen verzichtbar.

Aus methodischer Sicht ist die Verwendung einer Mehrjahresdatenbasis bei der statistischen Auffälligkeitseinstufung entscheidend, um die statistische Unsicherheit bei der Auslösung des Stellungnahmeverfahrens zu reduzieren. In Abschnitt 5.4 werden dazu mehrere, aus der statistischen Prozesskontrolle inspirierte Verfahren vorgeschlagen. Es wird gezeigt, dass diese Verfahren in einem theoretischen Setting eine bessere Trefferquote haben, als die 1-Jahres-Verfahren. Die entscheidende Frage ist, ob die Erweiterung der Datenbasis auch praktikable Lösungen für die Diskussionen im Stellungnahmeverfahren bietet. Klar ist, dass mit mehr Fällen besser Strukturen und Muster erkannt werden können. Wichtig wäre es somit, erste Erfahrungen in der Praxis mit einer solchen Vorgehensweise zu machen.

Fachliche Bewertung von Stellungnahmen

Ergibt sich aus einem Indikatorergebnis ein hinreichender Hinweis für ein Qualitätsdefizit gemäß der empfohlenen statistischen Methodik, soll dem Leistungserbringer die Gelegenheit eingeräumt werden, diesen Hinweis zu entkräften. Damit hat das Stellungnahmeverfahren die Funktion, das Vorliegen von Einflussfaktoren, die den Rückschluss vom Indikatorergebnis auf die Versorgungsqualität in Frage stellen können, aber noch nicht (genügend) in der

Berechnungsvorschrift des Indikators berücksichtigt sind, zu klären (Prüfung der Validität je Messung, vgl. Abschnitt 3.3). Diese Funktion von Stellungnahmen impliziert einen Bedeutungswandel des Stellungnahmeverfahrens weg von der umfassenden Bewertung der Versorgungsqualität hin zur Bewertung der Validität des Indikatorergebnisses im speziellen Fall. Durch die explizite Vorgabe von Entscheidungsregeln für diesen Prozess soll auch hier eine Steigerung der Objektivität des Verfahrens erreicht werden. Werden Qualitätsindikatoren als Aufgreifkriterien für einen nachfolgenden Peer-Review-Prozess verstanden, ist eine solche über das Qualitätsziel des Indikators hinausgehende Qualitätsbewertung sinnvoll. Im Verwendungskontext von *accountability* sind jedoch transparente und einheitliche Bewertungsregeln, wie sie hier vorgeschlagen werden, zentral (vgl. Abschnitt 3.2). Für die fachliche Bewertung der in den Stellungnahmen vorgebrachten Einflussfaktoren wird daher ein explizites Vorgehen einschließlich Nachberechnung des Indikatorergebnisses und der Vergleich mit einem (korrigierten) Referenzwert empfohlen. Damit sollen die Entscheidungsregeln, die zu einer Entkräftung der hinreichenden Evidenz für ein Qualitätsdefizit führen können, so transparent wie möglich gemacht werden. Auch aus wissenschaftlicher Sicht hat eine transparente Darstellung expliziter Entscheidungsregeln einen hohen Wert, da diese damit nachvollziehbar und kritisierbar werden. Eine solche Transparenz kann dadurch auch zu einer Steigerung der Akzeptanz des Verfahrens führen. Darüber hinaus soll die Akzeptanz des Verfahrens auch durch die fachliche Kompetenz, die Unabhängigkeit sowie die Interdisziplinarität der Fachkommissionen gesteigert werden. Kollegiale Gespräche und Begehungen (sowie die weiteren Förderinstrumente nach § 17 DeQS-RL) sollen im Sinne der Trennung von Qualitätsbewertung und -förderung zukünftig nicht mehr für die Ermittlung der Qualität der Versorgung eines Leistungserbringers eingesetzt werden. Dies steht im Einklang mit der Empfehlung, zukünftig nur noch beim Vorliegen überzeugender Gründe Indikatorergebnisse, die hinreichende Evidenz für ein Qualitätsdefizit nahelegen, in der fachlichen Bewertung mit „Kein Hinweis auf ein Qualitätsdefizit“ zu bewerten. Die überzeugende schriftliche Darlegung solcher Einflussfaktoren auf das Indikatorergebnis sollte Leistungserbringern, bei denen kein Qualitätsdefizit vorlag, auch ohne zusätzliche Begehung und kollegiales Gespräch möglich sein. Dies soll vor allem auch den Arbeitsaufwand bei der Bewertung der Stellungnahmen reduzieren.

Qualitätsförderung

Die Empfehlungen für qualitätsverbessernde Maßnahmen zielen darauf ab, bei den begrenzten Ressourcen der LAG, die Maßnahmen auf diejenigen Leistungserbringer zu fokussieren, die die größten Qualitätsdefizite und vermutlich auch das größte Verbesserungspotenzial aufweisen. Dabei soll jedoch den LAG bzw. Bundesstelle möglichst freie Hand bei der Wahl der Fördermaßnahme gelassen werden, da diese auf die individuelle Situation vor Ort angepasst werden muss. Denn im Gegensatz zu einem möglichst standardisierten und damit objektiven Vorgehen bei der Qualitätsbewertung ist für qualitätsverbessernde Maßnahmen ein möglichst auf den entsprechenden Leistungserbringer zugeschnittenes Vorgehen unter Berücksichtigung der in Abschnitt 3.7 beschriebenen Erfolgsfaktoren sinnvoll.

11.5.1 Überblicksliste der Empfehlungen

Im Folgenden werden die Empfehlungen stichpunktartig zusammengefasst:

Ermittlung einer statistischen Auffälligkeit und ihre Einordnung

- Die fallabhängigen Qualitätsindikatoren der DeQS-RL werden nach einer analytischen Herangehensweise berechnet und ausgewertet, d. h. durch stochastische Einflüsse entstehende Unsicherheit ist bei der Berechnung der Indikatorergebnisse und beim Abgleich mit dem Referenzbereich zu berücksichtigen.
- Der hinreichende Hinweis auf ein Qualitätsdefizit wird als vorgegebener Schwellenwert (entsprechend dem Signifikanzniveau) für die Posteriori-Verteilung der Null-Hypothese im Rahmen eines bayesianischen Hypothesentests statistisch operationalisiert.
- Der Auswertungs- und Bewertungsprozess der QI-Ergebnisse erfolgt jährlich.
- Ergibt sich anhand der vorliegenden Daten eines Leistungserbringers kein hinreichender Hinweis auf ein Qualitätsdefizit, wird das Indikatorergebnis mit „kein hinreichender Hinweis auf ein Qualitätsdefizit“ (U0) bewertet.
- Ergibt sich anhand der vorliegenen Daten eines Leistungserbringers ein hinreichender Hinweis auf ein Qualitätsdefizit, dann erhält der Leistungserbringer die Möglichkeit zur Stellungnahme.
- Das Signifikanzniveau α kann zunächst pro QS-Verfahren so gewählt werden, dass die Anzahl an anzufordernden Stellungnahmen konstant gegenüber der Anzahl der im vorangegangen Jahr angeforderten Stellungnahmen ist.
- Als Teil der Begleitung der Umsetzung trifft das IQTIG unter Aufwand-Nutzen-Überlegungen eine inhaltlich-methodisch begründete Festlegung des Signifikanzniveaus α. Berücksichtigt werden dabei die Güte der Indikatoren, der Aufwand und vorhandene Ressourcen für die Durchführung des Stellungnahmeverfahrens, Konsequenzen für Leistungserbringer und Patienten und Patientinnen bei Fehlklassifikationen, die Möglichkeit unterschiedlicher Signifikanzniveaus für unterschiedliche QI und/oder QS-Verfahren.
- Um eine höhere Treffsicherheit bei der statistischen Operationalisierung des hinreichenden Hinweises auf ein Qualitätsdefizit zu gewährleisten, soll perspektivisch eine Mehrjahresbetrachtung bei der Klassifikation der Leistungserbringerergebnisse stattfinden.

Formaler Ablauf des Stellungnahmeverfahrens

- Das Anforderungsschreiben für eine Stellungnahme soll am 30. Juni an die Leistungserbringer versendet werden.
- Der Leistungserbringer erhält ab Eingang des Anforderungsschreibens fünf Wochen Zeit, um eine qualifizierte Stellungnahme einzureichen.
- Erklärt der Leistungserbringer den Verzicht auf Abgabe einer Stellungnahme, wird das Indikatorergebnis mit „Qualitätsdefizit – Hinweis auf Qualitätsdefizit im Stellungnahmeverfahren bestätigt oder Verzicht auf Einreichung einer Stellungnahme“ (A3) bewertet.
- Geht die Stellungnahme nicht innerhalb der 5-wöchigen Frist ein, wird das Indikatorergebnis mit „Qualitätsdefizit – Stellungnahme nicht fristgerecht eingereicht oder Stellungnahme entsprach nicht den formalen Anforderungen“ (A0) bewertet.

- Stellungnahmen, die nicht den formalen Kriterien entsprechen, werden nicht berücksichtigt und das Indikatorergebnis wird mit „Qualitätsdefizit – Stellungnahme nicht fristgerecht eingereicht oder Stellungnahme entsprach nicht den formalen Anforderungen“ (A0) bewertet.
- Bis zum Ablauf der fünfwöchigen Frist kann der Leistungserbringer Korrekturen seiner Stellungnahme einreichen.

Formale Kriterien für Stellungnahmen

- Für jedes Indikatorergebnis, welches nach dem obigen statistischen Verfahren einen hinreichenden Hinweis auf ein Qualitätsdefizit liefert, ist eine Stellungnahme abzugeben.
- Die Stellungnahme einschließlich eventuell notwendiger Anhänge ist in Schriftform zu erstellen und einzureichen.
- Belege, die die Argumentation des Leistungserbringers unterstützen, können in angemessenem Umfang ergänzend zur Stellungnahme eingereicht werden, ersetzen diese jedoch nicht. Belege, die ohne erläuternden Text übermittelt werden, sollen nicht anerkannt werden.
- Die fachlich verantwortliche Instanz des Leistungserbringers (z. B. die Chefärztin/der Chefarzt, Praxisinhaberin/Praxisinhaber) ist über den Hinweis auf ein Qualitätsdefizit zu informieren und sie soll die angefertigte Stellungnahme autorisieren.
- Die Anonymität sowohl der Patientinnen und Patienten (personenidentifizierende Daten) als auch ggf. weiterer Beteiligter (z. B. Zuweiser) ist zu wahren. Bei Nichteinhaltung der Vorgaben zur Wahrung der Anonymität soll die Stellungnahme aus Datenschutzgründen vernichtet werden und der Leistungserbringer über diesen Vorgang benachrichtigt werden.
- Der Leistungserbringer soll eine Erklärung darüber abgeben, ob die übermittelten QS-Daten, auf denen das Indikatorergebnis basiert, aus seiner Sicht korrekt sind.
- Die Erfüllung der formalen Kriterien für Stellungnahmen prüft und dokumentiert die LAG bzw. die Bundesstelle bevor sie die Stellungnahme an die Fachkommission weiterleitet.

Inhaltliche Anforderungen an Stellungnahmen

- Der Leistungserbringer soll für jeden Grund, der aus seiner Sicht zur Abweichung vom Referenzbereich geführt hat, angeben, auf welche seiner Behandlungsfälle dieser Grund zutrifft bzw. ob der Grund für alle Behandlungsfälle zutrifft.
- Für die Zuordnung der Gründe zu den Behandlungsfällen sind die Vorgangsnummern der entsprechenden Behandlungsfälle mit der Stellungnahme zu übermitteln. Gleiches gilt auch für diejenigen Behandlungsfälle, bei denen kein interessierendes Ereignis (Behandlungsfälle, die nur im Nenner des Indikators vorkommen) aufgetreten ist.

Fachliche Bewertung

- Die Prüfung, ob Gründe für die statistische Auffälligkeit des Leistungserbringers vorliegen, die nicht vom Leistungserbringer zu vertreten sind und nicht schon in der Rechenregel des Indikators oder in dem Risikoadjustierungsmodell des Indikators berücksichtigt sind, heißt „fachliche Bewertung“ und wird von einer Fachkommission auf Basis der schriftlichen Stellungnahme des Leistungserbringers durchgeführt.

- Begehungen und Besprechungen bzw. kollegiale Gespräche werden ausschließlich im Rahmen der Qualitätsförderung eingesetzt.
- In der Stellungnahme angeführte Gründe, die aus Sicht des Leistungserbringers nicht von ihm zu verantworten waren, sollen (mittels schriftlichem Beleg) nachgewiesen werden.
- Die Fachkommission klassifiziert die angeführten Gründe dahingehend, ob sie vom Leistungserbringer zu verantworten sind oder nicht.
- Gründe, bei denen aus der Stellungnahme nicht hervorgeht, in welchem Maße sie auch bei Behandlungsfällen ohne interessierendes Ereignis vorgelegen haben, sollen in der fachlichen Bewertung nicht berücksichtigt werden.Die Fachkommission führt auf Basis der Klassifikation der angeführten Gründe mittels technischer Unterstützung eine Nachberechnung des Indikatorergebnisses für den Leistungserbringer durch. Dabei werden die Fälle, bei denen die Fachkommission zu der Einschätzung gelangt, dass die angeführten Gründe nicht vom Leistungserbringer zu verantworten sind, aus der Grundgesamtheit des Indikators ausgeschlossen.
- Liegt nach der Nachberechnung kein hinreichender statistischer Hinweis auf ein Qualitätsdefizit vor, dann erfolgt eine Bewertung des Indikatorergebnisses mit „Kein Hinweis auf Qualitätsdefizit“ (U).
- Liegt nach der Nachberechnung ein hinreichender statistischer Hinweis auf ein Qualitätsdefizit vor, prüft die Fachkommission, ob weitere, nicht eindeutig klassifizierbare Gründe vom Leistungserbringer geltend gemacht wurden.
- Wurden vom Leistungserbringer weitere, nicht eindeutig klassifizierbare Gründe geltend gemacht, beurteilt die Fachkommission auf Basis ihrer fachlichen Expertise, ob diese Gründe hinreichend sind, um die statistische Auffälligkeit nach der Nachberechnung zu erklären.
- Wenn die Fachkommission zu der Einschätzung gelangt, dass die angeführten Gründe hinreichend sind, um die statistische Auffälligkeit zu erklären, empfiehlt sie das Indikatorergebnis als „Kein Hinweis auf Qualitätsdefizit“ (U) zu bewerten.
- Wenn die Fachkommission zu der Einschätzung gelangt, dass die angeführten Gründe nicht hinreichend sind, um die statistische Auffälligkeit nach der Nachberechnung zu erklären, empfiehlt sie das Indikatorergebnis als „Qualitätsdefizit“ (A) zu bewerten.
- Die Fachkommission empfiehlt eine abschließende Bewertung des Indikatorergebnisses gemäß dem Bewertungsschema einschließlich der Einstufung in die Ziffern.
- Die LAG bzw. die Bundesstelle trifft die abschließende Entscheidung über die Einordnung des Indikatorergebnisses in das Bewertungsschema.
- Weicht die LAG bzw. die Bundesstelle von der Einstufungsempfehlung der Fachkommission ab, begründet und dokumentiert sie diese Abweichung.
- Mit der abschließenden Entscheidung in Bezug auf die Einordnung des Indikatorergebnisses in das Bewertungsschema ist die Qualitätsbewertung abgeschlossen.
- In Ausnahmefällen kann die LAG bzw. die Bundesstelle das Indikatorergebnis eines Leistungserbringers mit „Sonstiges – ohne Bewertung“ (S) bewerten. Der Sachverhalt, der eine Qualitätsbewertung unmöglich machte, ist durch die LAG bzw. die Bundesstelle anzugeben.

Zusammensetzung der Fachkommissionen

- Die LAG bzw. die Bundesstelle trägt die Verantwortung für die Auswahl der Expertinnen und Experten sowie die Zusammensetzung der Fachkommission.
- Die Fachkommissionen sollen interdisziplinär zusammengesetzt werden.
- In Abhängigkeit vom jeweiligen QS-Verfahren sind folgende Berufsgruppen einzubinden: Fachärztinnen und Fachärzte, Pflegefachkräfte, Klinische Qualitätsmanagerinnen und Qualitätsmanager, Patientenvertreterinnen und -vertreter, weitere Angehörige der Gesundheitsberufe, Krankenhaushygienikerinnen und -hygieniker oder Hygienefachkräfte

Qualitätsverbessernde Maßnahmen

- Qualitätsverbessernde Maßnahmen der LAG bzw. der Bundesstelle schließen sich nur an eine abgeschlossene Qualitätsbewertung an.
- Die Fachkommission empfiehlt unter Berücksichtigung des Indikatorergebnisses und seiner abschließenden Bewertung den Einsatz von Maßnahmen zur Qualitätsverbesserung.
- Die LAG bzw. die Bundesstelle leitet ab dem zweiten mit „Qualitätsdefizit“ (A) bewerteten Indikatorergebnis bei dem gleichen oder einem inhaltlich ähnlichen Indikator innerhalb von drei Jahren immer qualitätsverbessernde Maßnahmen ein.
- Es liegt im Ermessen der Fachkommission, unter Angabe von Gründen bereits bei nur einem mit „Qualitätsdefizit“ (A) bewerteten Indikatorergebnis qualitätsverbessernde Maßnahmen der LAG bzw. der Bundesstelle bei einem Leistungserbringer zu empfehlen.
- Die LAG bzw. Bundesstelle entscheidet über die Art und Einleitung von notwendigen Maßnahmen gemäß DeQS-RL und QFD-RL.
- Begehungen und Besprechungen bzw. kollegiale Gespräche werden der Maßnahmenstufe 1 zugeordnet.
- Der Einsatz und Umsetzungsgrad der qualitätsverbessernden Maßnahmen ist zu dokumentieren.

11.5.2 Fazit

Zusammenfassend würde eine Umsetzung der vorliegenden Empfehlungen deutliche Änderungen am bisherigen Vorgehen und an der DeQS-RL mit einer Annäherung an das Vorgehen nach plan. QI-RL bedeuten. Gleichzeitig wäre eine frühzeitigere Berichterstattung zur Versorgungsqualität möglich. Diese Änderungen müssten von verschiedenen Akteuren an verschiedenen Stellen umgesetzt werden. Während Änderungen in der statistischen Methodik für die Auswertung der Indikatorergebnisse im Rahmen der DeQS-RL vom IQTIG umgesetzt werden müssten, müssten die Änderungen, die sich für die Erstellung der Stellungnahmen ergeben, den Leistungserbringern erst umfassend kommuniziert werden. Die Nachberechnung im Rahmen der fachlichen Bewertungen durch die LAG und Fachkommission in Echtzeit setzt eine umfassende technische Unterstützung voraus, sodass die notwendige IT-seitige Entwicklungsarbeit bei einer möglichen Beauftragung der Umsetzung des Konzeptes berücksichtigt werden muss. Für die LAG, die Bundesstelle sowie die Fachkommissionen ergeben sich vermutlich die meisten Änderungen. Viele der empfohlenen Veränderungen zielen dabei auf eine Standardisierung des Vor-

gehens und auf eine Verschiebung der aufwendigen dialogischen Elemente von der Qualitätsbewertung hin zur Qualitätsförderung ab. Inwieweit sich die theoretisch zu erwartenden Effizienzsteigerungen in der praktischen Umsetzung realisieren lassen, lässt sich jedoch ohne erste praktische Erfahrungen mit den Empfehlungen nicht ohne Weiteres quantifizieren. Eine enge Abstimmung mit den LAG ist daher Voraussetzung für eine erfolgreiche Umsetzung. Die in Kapitel 10 gegebenen Empfehlungen zu einer formativen, also die Umsetzung begleitenden Evaluation spielen hier eine große Rolle. Beispielsweise könnte eine prozessbegleitende Befragung der LAG Hürden bei der Umsetzung der Empfehlungen zur Bewertung von Stellungnahmen erfragen und zu Verbesserungsvorschlägen führen.

Aufgrund der Vielzahl der Änderungen, die die Umsetzung des empfohlenen Konzepts mit sich bringen würde und aufgrund der Vielzahl der beteiligten Akteure, deren Handeln aufeinander abgestimmt werden müsste, empfiehlt das IQTIG eine strukturierte Begleitung der Umsetzung durch das IQTIG. Im Rahmen einer solchen Umsetzungsbegleitung sollte unter anderem ein Informationskonzept für die Leistungserbringer zu den geänderten Modalitäten des Stellungnahmeverfahrens umgesetzt, weitere Workshops mit den LAG-Vertreterinnen und -Vertreter sowie die formative Evaluation der Empfehlungen durchgeführt werden. Außerdem sollte in Abhängigkeit der umzusetzenden Empfehlungen eine schrittweise Umsetzung erfolgen und gegebenenfalls Übergangsregelungen getroffen werden.

Literatur

Adams, JL (2009): The Reliability of Provider Profiling. A Tutorial. Technical Report. [Stand:] 17.09.2019. Santa Monica, US-CA: RAND Corporation. Document Number: TR-653-NCQA. DOI: 10.7249/TR653.

Agresti, A (2013): Categorical Data Analysis. Third Edition. (Wiley Series in Probability and Statistics). Hoboken, US-NJ: Wiley. ISBN: 978-0-470-46363-5.

Altman, DG; Bland, JM (1995): Statistics Notes. Absence of evidence is not evidence of absence. *BMJ* 311:485. DOI: 10.1136/bmj.311.7003.485.

AQUA [Institut für angewandte Qualitätsförderung und Forschung im Gesundheitswesen] (2013): Bericht zum Strukturierten Dialog 2012. Erfassungsjahr 2011. Stand: 03.06.2013. Göttingen: AQUA. Signatur: 13-SQG-012. URL: https://sqg.de/upload/CONTENT/Themen/Strukturierter_Dialog/Bericht_Strukturierter_Dialog_2012.pdf (abgerufen am: 27.02.2019).

Ash, AS; Fienberg, SE; Louis, TA; Normand, SLT; Stukel, TA; Utts, J (2012): Statistical Issues in Assessing Hospital Performance. Commissioned by the Committee of Presidents of Statistical Societies [*White paper*]. Revised: 27.01.2012. Baltimore, US-MD: Centers for Medicare and Medicaid Services; Committee of Presidents of Statistical Societies. URL: https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/HospitalQualityInits/Downloads/Statistical-Issues-in-Assessing-Hospital-Performance.pdf (abgerufen am: 09.10.2019).

Åström, KJ (1965): Optimal Control of Markov Processes with Incomplete State Information. *Journal of Mathematical Analysis and Applications* 10(1): 174-205. DOI: 10.1016/0022-247x(65)90154-x.

BÄK [Bundesärztekammer] (2014): Leitfaden. Ärztliches Peer Review. (Texte und Materialien der Bundesärztekammer zur Fortbildung und Weiterbildung, Band 31). Berlin: BÄK. URL: https://www.bundesaerztekammer.de/fileadmin/user_upload/downloads/Leitfaden_Aerztliches-Peer-Review_2014.pdf (abgerufen am: 27.02.2019).

Berger, JO (2010): Statistical Decision Theory and Bayesian Analysis. Second Edition. (Springer Series in Statistics). New York, US-NY [u. a.]: Springer. ISBN: 978-1-4419-3074-3.

Berwick, DM (1991): Controlling Variation in Health Care: A Consultation from Walter Shewhart. *Medical Care* 29(12): 1212-1225.

Berwick, DM; James, B; Coye, MJ (2003): Connections Between Quality Measurement and Improvement. *Medical Care* 41(1 Suppl.): I–30-I–38. DOI: 10.1097/00005650-200301001-00004.

Brown, LD; Cai, TT; DasGupta, A (2001): Interval Estimation for a Binomial Proportion. *Statistical Science* 16(2): 101-133. DOI: 10.1214/ss/1009213286.

Carlin, BP; Louis, TA (2008): Bayesian Methods for Data Analysis Third Edition. (Texts in Statistical Science). Boca Raton, US-FL [u. a.]: Chapman & Hall/CRC. ISBN: 978-1-58488-697-6.

Carter, N (1989): Performance Indicators: 'backseat driving' or 'hands off' control? *Policy and Politics* 17(2): 131-138. DOI: 10.1332/030557389782454857.

Chang, W; Cheng, J; Allaire, JJ; Xie, Y; McPherson, J (2019): shiny: Web Application Framework for R [*Open Source Software*]. Version 1.3.2. Published: 22.04.2019. Vienna, AT: R Foundation for Statistical Computing. URL: https://CRAN.R-project.org/package=shiny [Downloads: shiny_1.3.2.tar.gz] (abgerufen am: 09.10.2019).

Chassin, MR; Loeb, JM; Schmaltz, SP; Wachter, RM (2010): Accountability Measures – Using Measurement to Promote Quality Improvement. *The New England Journal of Medicine* 363(7): 683-688. DOI: 10.1056/NEJMsb1002320.

Chop, I; Eberlein-Gonska, M (2012): Übersichtsartikel zum Peer Review Verfahren und seine Einordnung in der Medizin. *Zeitschrift für Evidenz, Fortbildung und Qualität im Gesundheitswesen* 106(8): 547-552. DOI: 10.1016/j.zefq.2012.08.017.

Christiansen, CL; Morris, CN (1996): Fitting and Checking a Two-Level Poisson Model: Modeling Patient Mortality Rates in Heart Transplant Patients. Kapitel 18. In: Berry, DA; Stangl, D; Hrsg.: *Bayesian Biostatistics.* (Statistics: textbooks and monographs, volume 151). Boca Raton, US-FL [u. a.]: Chapman & Hall/CRC, 467-501. ISBN: 0-8247-9334-X.

Dawes, RM; Faust, D; Meehl, PE (1989): Clinical Versus Actuarial Judgment. *Science* 243(4899): 1668-1674. DOI: 10.1126/science.2648573.

DeGEval [Gesellschaft für Evaluation] (2008): Standards für Evaluation. In: DeGEval; Hrsg.: *Standards für Evaluation.* 4. unveränderte Auflage. [Stand:] Juli 2008. Mainz: DeGEval, 10-13. ISBN: 3-00-009022-3. URL: https://www.degeval.org/fileadmin/user_upload/Sonstiges/STANDARDS_2008-12.pdf (abgerufen am: 09.10.2019).

Deming, WE (1953): On the Distinction between Enumerative and Analytic Surveys. *Journal of the American Statistical Association* 48(262): 244-255. DOI: 10.1080/01621459.1953.10483470.

Dimick, JB; Staiger, DO; Birkmeyer, JD (2010): Ranking Hospitals on Surgical Mortality: The Importance of Reliability Adjustment. *Health Services Research* 45: 1614-1629. DOI: 10.1111/j.1475-6773.2010.01158.x.

Djulbegovic, B; Beckstead, JW; Elqayam, S; Reljic, T; Hozo, I; Kumar, A; et al. (2014): Evaluation of Physicians' Cognitive Styles. *Medical Decision Making* 34(5): 627-637. DOI: 10.1177/0272989x14525855.

Donabedian, A (2003): An Introduction to Quality Assurance in Health Care. Oxford, GB [u. a.]: Oxford University Press. ISBN: 978-0-19-515809-0.

Eberlein-Gonska, M; Rink, O; Winklmair, C (2017): Wie wir Qualität verbessern. Kapitel 5. In: Eberlein-Gonska, M; Martin, J; Zacher, J; Hrsg.: *Handbuch IQM. Konsequent transparent –*

Qualität mit Routinedaten. 2. aktualisierte und erweiterte Auflage. Berlin: Medizinisch Wissenschaftliche Verlagsgesellschaft, 53-65. ISBN: 978-3-95466-115-2. URL: https://www.initiative-qualitaetsmedizin.de/mediapool/1736/media_file/sendfile/ (abgerufen am: 09.10.2019).

Freeman, T (2002): Using performance indicators to improve health care quality in the public sector: a review of the literature. *Health Services Management Research* 15(2): 126-137. DOI: 10.1258/0951484021912897.

G-BA [Gemeinsamer Bundesausschuss] (2018a): Beschluss des Gemeinsamen Bundesausschusses über eine Beauftragung des IQTIG: Weiterentwicklung des Strukturierten Dialogs mit Krankenhäusern. [Stand:] 18.01.2018. Berlin: G-BA. URL: https://www.g-ba.de/downloads/39-261-3196/2018-01-18_IQTIG-Beauftragung_Weiterentwicklung-strukturierter-Dialog.pdf (abgerufen am: 27.02.2019).

G-BA [Gemeinsamer Bundesausschuss] (2018b): Beschluss des Gemeinsamen Bundesausschusses über eine Richtlinie zur datengestützten einrichtungsübergreifenden Qualitätssicherung. [Stand:] 19.07.2018. Berlin: G-BA. BAnz AT 18.12.2018 B3. URL: https://www.g-ba.de/downloads/39-261-3419/2018-07-19_DeQS-RL_Erstfassung.pdf (abgerufen am: 28.11.2018).

G-BA [Gemeinsamer Bundesausschuss] (2019): Beschluss des Gemeinsamen Bundesausschusses über eine Beauftragung des IQTIG mit einer Ursachenanalyse der Auffälligkeiten sowie der methodischen Weiterentwicklung der Qualitätsindikatoren zur präoperativen Verweildauer bei der Versorgung der hüftgelenknahen Femurfraktur. [Stand:] 17.01.2019. Berlin: G-BA. URL: https://www.g-ba.de/downloads/39-261-3666/2019-01-17_IQTIG-Beauftragung_QS-Indikatoren-h%C3%BCftgelenknahe-Femurfraktur.pdf (abgerufen am: 10.10.2019).

Gardner, K; Olney, S; Dickinson, H (2018): Getting smarter with data: understanding tensions in the use of data in assurance and improvement-oriented performance management systems to improve their implementation. *Health Research Policy and Systems* 16:125. DOI: 10.1186/s12961-018-0401-2.

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

Gerber-Grote, A; Sandmann, FG; Zhou, M; ten Thoren, C; Schwalm, A; Weigel, C; et al. (2014): Decision making in Germany: Is health economic evaluation as a supporting tool a sleeping beauty? *Zeitschrift für Evidenz, Fortbildung und Qualität im Gesundheitswesen* 108(7): 390-396. DOI: 10.1016/j.zefq.2014.06.018.

Gerlach, F (2001): Konsequenzen: Empfehlungen für die eigene Tätigkeit in Praxis und Klinik. Kapitel 8. In: Gerlach, F: *Qualitätsförderung in Praxis und Klinik. Eine Chance für die Medizin.* Stuttgart [u. a.]: Georg Thieme, 267-285. ISBN: 978-3-13-125891-5.

Griem, C; Kleudgen, S; Diel, F (2013): Instrumente der kollegialen Qualitätsförderung. *Deutsches Ärzteblatt* 110(26): A1310-A1313, A5. URL: https://www.aerzteblatt.de/pdf.asp?id=141971 (abgerufen am: 09.10.2019).

Hengelbrock, J; Höhle, M (2019): Evaluating quality of hospital care using time-to-event endpoints based on patient follow-up data. *Health Services and Outcomes Research Methodology*, Epub 08.07.2019. DOI: 10.1007/s10742-019-00202-7.

Hensen, P (2016): Qualitätsmanagement im Gesundheitswesen. Grundlagen für Studium und Praxis. Wiesbaden: Springer Gabler. ISBN: 978-3-658-07744-0.

Höhle, M; Paul, M (2008): Count data regression charts for the monitoring of surveillance time series. *Computational Statistics & Data Analysis* 52(9): 4357-4368. DOI: 10.1016/j.csda.2008.02.015.

Huang, W; Reynolds, MR; Wang, S (2012): A Binomial GLR Control Chart for Monitoring a Proportion. *Journal of Quality Technology* 44(3): 192-208. DOI: 10.1080/00224065.2012.11917895.

Hudson, DW; Holzmueller, CG; Pronovost, PJ; Gianci, SJ; Pate, ZT; Wahr, J; et al. (2012): Toward Improving Patient Safety Through Voluntary Peer-to-Peer Assessment. *American Journal of Medical Quality* 27(3): 201-209. DOI: 10.1177/1062860611421981.

IQTIG [Institut für Qualitätssicherung und Transparenz im Gesundheitswesen] (2016): Planungsrelevante Qualitätsindikatoren. Abschlussbericht zur Auswahl und Umsetzung. Stand: 31.08.2016. Berlin: IQTIG. URL: https://iqtig.org/downloads/berichte/2016/IQTIG_Planungsrelevante-Qualitaetsindikatoren_Abschlussbericht.pdf (abgerufen am: 09.10.2019).

IQTIG [Institut für Qualitätssicherung und Transparenz im Gesundheitswesen] (2017a): Bericht zum Strukturierten Dialog 2016. Erfassungsjahr 2015. Stand: 15.05.2017. Berlin: IQTIG. URL: https://iqtig.org/downloads/berichte/2015/IQTIG_Bericht-zum-Strukturierten-Dialog-2016.pdf (abgerufen am: 27.02.2019).

IQTIG [Institut für Qualitätssicherung und Transparenz im Gesundheitswesen] (2017b): Bericht zum Strukturierten Dialog 2016. Erfassungsjahr 2015. Anhang. Stand: 15.05.2017. Berlin: IQTIG. URL: https://iqtig.org/downloads/berichte/2015/IQTIG_Bericht-zum-Strukturierten-Dialog-2016_Anhang.pdf (abgerufen am: 27.02.2019).

IQTIG [Institut für Qualitätssicherung und Transparenz im Gesundheitswesen] (2017c): Ereigniszeitanalyse-Methodik für die Follow-up-Indikatoren nach QSKH-RL. Stand: 06.04.2017. Berlin: IQTIG. URL: https://iqtig.org/dateien/berichte/2017/IQTIG_Ereigniszeitanalyse-Methodik-fuer-Follow-up-Indikatoren-nach-QSKH-RL_2017-04-06.pdf (abgerufen am: 23.09.2019).

IQTIG [Institut für Qualitätssicherung und Transparenz im Gesundheitswesen] (2017d): Methodische Grundlagen V1.0. Stand: 15.09.2017. Berlin: IQTIG. URL: https://iqtig.org/downloads/berichte/2017/IQTIG_Methodische-Grundlagen-V1.0.pdf (abgerufen am: 15.05.2018).

IQTIG [Institut für Qualitätssicherung und Transparenz im Gesundheitswesen] (2018a): Bericht zum Strukturierten Dialog 2017. Erfassungsjahr 2016. Stand: 24.08.2018. Berlin: IQTIG. URL: https://iqtig.org/downloads/berichte/2016/IQTIG_Bericht-zum-Strukturierten-Dialog-2017_2018-08-24_barrierefrei.pdf (abgerufen am: 09.10.2019).

IQTIG [Institut für Qualitätssicherung und Transparenz im Gesundheitswesen] (2018b): Biometrische Methodik für die Follow-up-Indikatoren nach QSKH-RL für die Bundesauswertung des Auswertungsjahres 2017. Stand: 23.10.2018. Berlin: IQTIG. URL: https://iqtig.org/downloads/berichte/2017/IQTIG_Biometrische-Methodik-FU-Indikatoren_BuAW-AJ-2017_2018-10-23.pdf (abgerufen am: 23.09.2019).

IQTIG [Institut für Qualitätssicherung und Transparenz im Gesundheitswesen] (2018c): Entwicklung einer Befragung von Patienten und Patientinnen mit Schizophrenie, schizotypen und wahnhaften Störungen. Entwicklung einer Patientenbefragung im Rahmen der Aktualisierung und Erweiterung des QS-Verfahrens *Versorgung von volljährigen Patienten und Patientinnen mit Schizophrenie, schizotypen und wahnhaften Störungen.* Abschlussbericht. Stand: 15.12.2018. Berlin: IQTIG. URL: https://iqtig.org/downloads/berichte/2018/IQTIG_Patientenbefragung_QS-Verfahren-Schizophrenie_Abschlussbericht-mit-AT_2018-12-15.pdf (abgerufen am: 29.01.2020).

IQTIG [Institut für Qualitätssicherung und Transparenz im Gesundheitswesen] (2018d): Entwicklung von Patientenbefragungen im Rahmen des Qualitätssicherungsverfahrens *Perkutane Koronarintervention und Koronarangiographie.* Abschlussbericht. Stand: 15.12.2018. Berlin: IQTIG. URL: https://iqtig.org/downloads/berichte/2018/IQTIG_Patientenbefragung_QS-PCI_Abschlussbericht-mit-AT_2018-12-15.pdf (abgerufen am: 20.01.2020).

IQTIG [Institut für Qualitätssicherung und Transparenz im Gesundheitswesen] (2018e): Hüftendoprothesenversorgung. Beschreibung der Qualitätsindikatoren für das Erfassungsjahr 2017. Indikatoren 2017. Stand: 06.04.2018. Berlin: IQTIG. URL: https://iqtig.org/downloads/auswertung/2017/hep/QSKH_HEP_2017_QIDB_V01_2018-04-06.pdf (abgerufen am: 26.06.2019).

IQTIG [Institut für Qualitätssicherung und Transparenz im Gesundheitswesen] (2018f): Kriterien für den gezielten Datenabgleich in der Datenvalidierung nach QSKH-RL. Abschlussbericht. Stand: 19.12.2018. Berlin: IQTIG. [unveröffentlicht].

IQTIG [Institut für Qualitätssicherung und Transparenz im Gesundheitswesen] (2019a): 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: 08.05.2019).

IQTIG [Institut für Qualitätssicherung und Transparenz im Gesundheitswesen] (2019b): Methodische Grundlagen V1.1s. Würdigung der Stellungnahmen. Stand: 15.04.2019. Berlin: IQTIG. URL: https://iqtig.org/dateien/dasiqtig/grundlagen/IQTIG_Methodische-Grundlagen-V1-1s_Wuerdigung-der-Stellungnahmen_2019-04-15.pdf (abgerufen am: 12.09.2019).

Jensen, FV; Nielsen, TD (2007): Bayesian Networks and Decision Graphs. Second Edition. (Information Science & Statistics). Berlin [u. a.]: Springer. ISBN: 978-0-387-68281-5.

Kaelbling, LP; Littman, ML; Cassandra, AR (1998): Planning and acting in partially observable stochastic domains. *Artificial Intelligence* 101(1-2): 99-134. DOI: 10.1016/s0004-3702(98)00023-x.

Kauermann, G; Küchenhoff, H (2011): Stichproben: Methoden und praktische Umsetzung mit R. Berlin [u. a.]: Springer. ISBN: 978-3-642-12317-7.

Keiding, N; Clayton, D (2014): Standardization and Control for Confounding in Observational Studies: A Historical Perspective. *Statistical Science* 29(4): 529-558. DOI: 10.1214/13-STS453.

Krell, RW; Hozain, A; Kao, LS; Dimick, JB (2014a): Reliability of Risk-Adjusted Outcomes for Profiling Hospital Surgical Quality. *JAMA Surgery* 149(5): 467-474. DOI: 10.1001/jamasurg.2013.4249.

Krell, RW; Staiger, DO; Dimick, JB (2014b): Reliability of Surgical Outcomes for Predicting Future Hospital Performance. *Medical Care* 52(6): 565-571. DOI: 10.1097/mlr.0000000000000138.

Krohne, HW; Hock, M (2015): Psychologische Diagnostik. Grundlagen und Anwendungsfelder. 2., überarbeitete und aktualisierte Auflage. Stuttgart: Kohlhammer. ISBN: 978-3-17-025255-4.

Kruglanski, AW; Gigerenzer, G (2011): Intuitive and Deliberate Judgments Are Based on Common Principles. *Psychological Review* 118(1): 97-109. DOI: 10.1037/a0020762.

Kuncel, NR; Klieger, DM; Connelly, BS; Ones, DS (2013): Mechanical Versus Clinical Data Combination in Selection and Admissions Decisions: A Meta-Analysis. *Journal of Applied Psychology* 98(6): 1060-1072. DOI: 10.1037/a0034156.

Lauritzen, SL; Nilsson, D (2001): Representing and Solving Decision Problems with Limited Information. *Management Science* 47(9): 1235-1251. DOI: 10.1287/mnsc.47.9.1235.9779.

Lieb, K; Klemperer, D; Koch, K; Baethge, C; Ollenschläger, G; Ludwig, WD (2011): Interessenkonflikte in der Medizin. Mit Transparenz Vertrauen stärken. *Deutsches Ärzteblatt* 108(6): A256-A260. URL: http://www.aerzteblatt.de/int/article.asp?id=80790 [PDF-Version > Download] (abgerufen am: 30.09.2019).

Liu, J; Louis, TA; Pan, W; Ma, JZ; Collins, AJ (2003): Methods for Estimating and Interpreting Provider-Specific Standardized Mortality Ratios. *Health Services and Outcomes Research Methodology* 4(3): 135-149. DOI: 10.1023/B:HSOR.0000031400.77979.b6.

Montgomery, DC (2013): Statistical Quality Control. A Modern Introduction. Seventh Edition. Hoboken, US-NJ: Wiley. ISBN: 978-1-118-32257-4.

Morris, CN (1983): Parametric Empirical Bayes Inference: Theory and Applications. *Journal of the American Statistical Association* 78(381): 47-55. DOI: 10.2307/2287098.

Nimptsch, U; Peschke, D; Mansky, T (2016): Der Einfluss von Qualitätsmessung, Transparenz und Peer Reviews auf die Krankenhaussterblichkeit – Retrospektive Vorher-Nachher-Studie mit 63 Kliniken. *Zeitschrift für Evidenz, Fortbildung und Qualität im Gesundheitswesen* 115-116: 10-23. DOI: 10.1016/j.zefq.2016.05.007.

Paddock, SM (2014): Statistical Benchmarks for Health Care Provider Performance Assessment: A Comparison of Standard Approaches to a Hierarchical Bayesian Histogram-Based Method. *Health Services Research* 49(3): 1056-1073. DOI: 10.1111/1475-6773.12149.

Pearl, J (2009): Causality: Models, Reasoning, and Inference. Second Edition. Cambridge, GB [u. a.]: Cambridge University Press. ISBN: 978-0-521-89560-6.

Profit, J; Typpo, KV; Hysong, SJ; Woodard, LD; Kallen, MA; Petersen, LA (2010): Improving benchmarking by using an explicit framework for the development of composite indicators: an example using pediatric quality of care. *Implementation Science* 5:13. DOI: 10.1186/1748-5908-5-13.

Raats, VM; Moors, JJA (2003): Double-checking auditors: a Bayesian approach. *Journal of the Royal Statistical Society. Series D (The Statistician)* 52(3): 351-365. DOI: 10.1111/1467-9884.00364.

Rink, O (2013): Das IQM-Peer-Review-Verfahren. Verbesserung der Ergebnisqualität, Strategie und Ergebnisse. *Prävention und Gesundheitsförderung* 8(1): 22-28. DOI: 10.1007/s11553-012-0374-x.

Schmidt-Atzert, L; Amelang, M (2012): Psychologische Diagnostik. 5. Auflage. Berlin [u. a.]: Springer. ISBN: 978-3-642-17000-3.

Shenoy, PP (1992): Valuation-Based Systems for Bayesian Decision Analysis. *Operations Research* 40(3): 463-484. DOI: 10.1287/opre.40.3.463.

Shwartz, M; Cohen, AB; Restuccia, JD; Ren, ZJ; Labonte, A; Theokary, C; et al. (2011): How Well Can We Identify the High-Performing Hospital? *Medical Care Research and Review* 68(3): 290-310. DOI: 10.1177/1077558710386115.

Shwartz, M; Restuccia, JD; Rosen, AK (2015): Composite Measures of Health Care Provider Performance: A Description of Approaches. *The Milbank Quarterly* 93(4): 788-825. DOI: 10.1111/1468-0009.12165.

Solberg, LI; Mosser, G; McDonald, S (1997): The Three Faces of Performance Measurement: Improvement, Accountability, and Research. *The Joint Commission Journal on Quality Improvement* 23(3): 135-147. DOI: 10.1016/S1070-3241(16)30305-4.

Spiegelhalter, D; Sherlaw-Johnson, C; Bardsley, M; Blunt, I; Wood, C; Grigg, O (2012): Statistical methods for healthcare regulation: rating, screening and surveillance. *Journal of the Royal Statistical Society. Series A (Statistics in Society)* 175(1): 1-47. DOI: 10.1111/j.1467-985X.2011.01010.x.

Spiegelhalter, DJ (2005): Handling over-dispersion of performance indicators. *Quality and Safety in Health Care* 14(5): 347-351. DOI: 10.1136/qshc.2005.013755.

Stroup, WW (2013): Generalized Linear Mixed Models. Modern Concepts, Methods and Applications. (Texts in Statistical Science). Boca Raton, US-FL [u. a.]: Chapman & Hall/CRC. ISBN: 978-1-4398-1512-0.

Tartakovsky, A; Nikiforov, I; Basseville, M (2015): Sequential Analysis. Hypothesis Testing and Changepoint Detection. (Monographs on Statistics and Applied Probability 136). Boca Raton, US-FL [u. a.]: Chapman & Hall/CRC. ISBN: 978-1-4398-3820-4.

Tenenbein, A (1970): A Double Sampling Scheme for Estimating from Binomial Data with Misclassifications. *Journal of the American Statistical Association* 65(331): 1350-1361. DOI: 10.2307/2284301.

Veit, C; Lüken, F; Bungard, S; Trümner, A; Tewes, C; Hertle, D (2013): Rahmenkonzept Evaluation bezogen auf Evaluationen nach § 137b SGB V. Version 1.1. Entwurf vom 17.07.2013. Düsseldorf: BQS [Institut für Qualität & Patientensicherheit]. [unveröffentlicht].

W. K. Kellogg Foundation (2004): Logic Model Development Guide. Using Logic Models to Bring Together Planning, Evaluation, and Action. Updated January 2004. Battle Creek, US-MI: W. K. Kellogg Foundation. URL: https://www.wkkf.org/resource-directory/resource/2006/02/wk-kellogg-foundation-logic-model-development-guide [Download PDF - Angabe von Name und E-Mail erforderlich] (abgerufen am: 09.10.2019).

Westen, D; Weinberger, J (2004): When Clinical Description Becomes Statistical Prediction. *American Psychologist* 59(7): 595-613. DOI: 10.1037/0003-066X.59.7.595.

Winkler-Komp, G; Misselwitz, B; Kupfernagel, F; van Emmerich, C; Döbler, K (2014): Maßnahmen zur Qualitätssicherung in Krankenhäusern: Strukturierter Dialog – Strukturen und Prozesse. Ergebnis einer Umfrage des Gemeinsamen Bundesausschusses bei den auf Landesebene beauftragten Stellen und der Institution nach § 137 a SGB V. *Das Krankenhaus* 2014(3): 198-205.

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.