Winterthurer Institut für Gesundheitsökonomie

Zürcher Hochschule für Angewandte Wissenschaften

Gewichtung im Rahmen des Spital Benchmarkings

Die vorliegende Studie wurde erstellt im Auftrag der Einkaufsgemeinschaft HSK von

Beatrice Brunner, Janina Nemitz und Matthias Maurer

Winterthur, 01. April 2020

Korrespondenz:

Beatrice Brunner
Co-Leiterin Gesundheitsökonomische Forschung
WIG / ZHAW
Gertrudstrasse 15, 8401 Winterthur
058 934 46 04
beatrice.brunner@zhaw.ch

Inhaltsverzeichnis

Zusammenfassung

Hintergrund

Seit der Einführung der neuen Spitalfinanzierung im Jahr 2012 gilt ein kostenbasiertes «Preissystem», wonach sich die Spitaltarife an der Entschädigung jener Spitäler zu orientieren haben, welche die tarifierte Leistung in der notwendigen Qualität effizient und günstig erbringen (Artikel 49 Abs. 1 KVG). Der Bundesrat ordnet dafür schweizweite Betriebsvergleiche zwischen Spitälern an. Bis heute fehlen jedoch hoheitliche Vorgaben zur Art und Weise, wie diese Betriebsvergleiche zu erfolgen haben.

In der Praxis erfolgen die Betriebsvergleiche anhand eines Benchmarkings, welches auf den Fallnormkosten (durchschnittliche schweregradbereinigte Fallkosten) der Spitäler basiert. Zur Ermittlung des «effizienten Preises» wird die Perzentil-Methode angewendet. Diese wurde gemäss Gerichtsentscheid als akzeptables Vergleichsverfahren anerkannt [1]. Gemäss dieser Methode wird der «effiziente Preis» (Benchmark) bei demjenigen Spital angesetzt, welches in der Reihenfolge der Fallnormkosten dem Perzentil X entspricht. Alle darunter liegenden Spitäler gelten als effizient, die darüber liegenden als ineffizient.

Uneinigkeit besteht in drei Aspekten: Erstens hinsichtlich der Höhe des angemessenen Perzentil-Werts, welcher für die Festlegung des Benchmarks massgebend sein soll. Zweitens wird die Sachgerechtigkeit der Fallnormkosten als Grundlage für das Benchmarking angezweifelt. Und drittens wird kontrovers diskutiert, ob das Benchmarking gewichtet (z.B. nach Anzahl Fällen oder Case Mix) oder ungewichtet vorgenommen werden soll [1], [2]. Oder, ob alternative Korrekturen, wie z.B. das von der HSK verwendete «Clustering», im Rahmen des Benchmarkings erfolgen sollen [3], [4].

Ziel und Methode

Die vorliegende Studie erarbeitet die Motive für und beurteilt die Zweckmässigkeit von einer Gewichtung im Rahmen des Spital Benchmarkings. Dazu werden folgende Fragen beantwortet:

1. Wann und zu welchem Zweck wird in der angewandten Statistik mit Gewichten gerechnet? Und welche Vor- und Nachteile entstehen dadurch?
2. Wann und in welcher Form ist eine Gewichtung im Rahmen des Spital Benchmarkings sinnvoll?

Zur Beantwortung der zweiten Frage wurde insbesondere auch die Sachgerechtigkeit der Fallnormkosten als Grundlage für das Spital Benchmarking diskutiert und beurteilt.

Methodisch basiert die Studie auf einer Literaturrecherche und eigenen konzeptionellen Überlegungen. Zur Verdeutlichung der Schlussfolgerungen wurde ein fiktives Benchmarking-Beispiel für 10 Spitäler entwickelt.

Resultate

Die Studie kommt zu vier zentralen Ergebnissen:

1. Es gibt derzeit verschiedene Ursachen für eine Verzerrung der Fallnormkosten, die eine Gewichtung denkbar machen. Diese Ursachen unterscheiden sich jedoch hinsichtlich ihrer Relevanz. Verzerrungen, die auf eine unvollständige Datengrundlage oder auf die inkorrekte und uneinheitliche Berechnung der benchmarkrelevanten Betriebskosten zurückzuführen sind, lassen sich langfristig beheben (z.B. durch vollständigere Daten und methodische Verbesserungen) und werden deshalb als weniger relevant eingestuft. Verzerrungen, die hingegen auf Fehlbewertungen von DRG Kostengewichten (z.B. wegen kleiner Fallgruppen oder Outlier-Korrektu-

ren) und Kosteninhomogenitäten innerhalb von DRGs (z.B. wegen leistungs- und patientenbezogenen Unterschieden zwischen den Spitälern) zurückzuführen sind, lassen sich nicht vollständig im Sinne einer Evolution der heutigen Tarifstruktur beheben. Sie sind system-inhärent und daher auch langfristig von Bedeutung.

2. Wenn sich die Ursache für die Verzerrung anderweitig beheben lässt, ist im Rahmen des Spital Benchmarkings auf jeden Fall von Gewichten abzuraten. Der Grund ist, dass ausschliesslich modellbasierte Gewichte zur Anwendung kämen. Diese Gewichte müssen geschätzt werden, basierend auf sehr starken Annahmen, die nicht überprüft werden können. So ist es grundsätzlich möglich, dass diese Gewichte mehr Schaden anrichten, als dass sie beheben. Aus diesem Grund steht man der Anwendung solcher Gewichte kritisch gegenüber. Ein begründetes Gewichtungsmotiv geht somit nur von zwei Verzerrungsursachen aus: Fehlbewertungen von DRG Kostengewichten und Kosteninhomogenitäten innerhalb von DRGs.

3. Die Anwendung von Gewichten im Rahmen des Spital Benchmarkings ist in keinem Fall zweckmässig. Zwar könnte bei Verwendung von korrekten Gewichten der Benchmark-Wert korrekt berechnet werden. Eine Unterteilung in effiziente und ineffiziente Spitäler ist aber dennoch nicht möglich, weil sich durch die Gewichtung der Spitäler die Fallnormkosten und somit die Reihenfolge der Spitäler nicht verändern.

4. Eine korrekte Unterteilung in effiziente und ineffiziente Spitäler kann ausschliesslich über eine Korrektur der Fallnormkosten erreicht werden. Die Korrektur der Fallnormkosten erfolgt dabei idealerweise anhand eines empirischen Modells, wie z.B. dem «Fallpauschalenmodell» von Polynomics [5]. Die auf diese Weise um die gerechtfertigten Unterschiede korrigierten Fallnormkosten erlauben ein valides schweizweites Benchmarking. Es müssten dann zwar immer noch differenzierte Baserates verhandelt werden, weil die Korrektur nachgelagert an die Berechnung der Kostengewichte durch SwissDRG erfolgt. Die Höhe der Verzerrung der Fallnormkosten würde aber eine valide Grundlage für die Verhandlungen bilden. Alternativ wäre eine Art «gerechtfertigter» Lastenausgleich zwischen den Spitälern denkbar, der nachträglich für die Verzerrungen in den Fallnormkosten kompensiert. Dies hätte den Vorteil, dass es nur noch eine schweizweit einheitlich geltende Baserate gäbe und somit keine Verhandlungen mehr nötig wären. Auch diese Lösung setzt die Korrektur (Sachgerechtigkeit) der Fallnormkosten voraus, um die Höhe etwaiger Ausgleichszahlungen zwischen den Spitälern ex-post zu bestimmen.

1 Ausgangslage, Ziel, Methodik

1.1 Hintergrund

Mit der Einführung der neuen Spitalfinanzierung gilt seit 2012 ein kostenbasiertes «Preissystem». So haben sich nach Artikel 49 Abs. 1 KVG die Spitaltarife an der Entschädigung jener Spitäler zu orientieren, welche die tarifierte Leistung in der notwendigen Qualität effizient und günstig erbringen. Dafür ordnet der Bundesrat schweizweite Betriebsvergleiche zwischen Spitälern an (Artikel 49 Abs. 8 KVG). Zur Art und Weise wie diese Betriebsvergleiche zu erfolgen haben, hat der Bundesrat bis heute jedoch noch keine detaillierten Vorgaben erlassen. Solche sind für das Jahr 2020 angekündigt [6].

Die bisher bestehenden Vorgaben sind mehrheitlich das Resultat von Tarifstreitigkeiten vor dem Bundesverwaltungsgericht. Dieses hat u.a. Vorgaben für die einheitliche Herleitung der stationären für den Spitalvergleich relevanten Betriebskosten festgelegt. Ausserdem wurde gemäss Zürcher und Glarner Grundsatzurteilen das Benchmarking mittels Perzentil-Methode als akzeptables Vergleichsverfahren anerkannt [2]. Gemäss dieser Methode wird der Referenzwert (Benchmark) bei dem Spital angesetzt, welches in der Reihenfolge der Fallnormkosten (schweregradbereinigte durchschnittliche Fallkosten) dem Perzentil X entspricht. Alle unter dem Benchmark liegenden Spitäler gelten als effizient, die darüber liegenden als ineffizient. Uneinig ist man sich bis heute über die Höhe eines angemessenen Perzentil-Werts, welcher für die Festlegung des Benchmarks massgebend ist. Darüber hinaus wird die Frage, ob die Kalkulation gewichtet (z.B. nach Anzahl Fällen oder Case Mix) oder ungewichtet vorgenommen werden soll, kontrovers diskutiert [1], [2].

1.2 Ziele und Fragestellungen

Ziel dieser Studie ist die konzeptionelle Erarbeitung und Diskussion der Gründe, die für bzw. gegen eine Gewichtung im Rahmen des Spital Benchmarkings sprechen. Zur Zielerreichung dienen uns zwei Fragestellungen:

1. Wann und zu welchem Zweck wird in der angewandten Statistik mit Gewichten gerechnet? Und welche Vor- und Nachteile entstehen dadurch?
2. Wann und in welcher Form ist eine Gewichtung im Rahmen des Spital Benchmarkings sinnvoll?
   2.1. Was ist das Ziel des Spital Benchmarkings?
   2.2. Wie valide sind die Fallnormkosten als Grundlage für das Spital Benchmarking?
   2.3. Was hat Gewichtung für Auswirkungen auf das Resultat des Spital Benchmarkings?

Die Studie grenzt sich im Besonderen von drei Themen ab:

Erstens: Die Grundsatzfrage, welches Perzentil angemessen ist, wird in dieser Studie nicht diskutiert. Grundsätzlich handelt es sich bei dieser Frage um eine normative Vorgabe darüber, wie intensiv der Wettbewerb ausfallen soll. Den Tarifparteien und Kantonen ist bisher ein weiter Beurteilungs- und Ermessensspielraum eingeräumt worden. Wir gehen in unserem Bericht vom 25ten Perzentil aus.

Zweitens: Die Berechnung der benchmarkrelevanten Betriebskosten auf Basis von ITAR-K Daten wird nicht hinterfragt und beurteilt. Hierzu sind mit dem vom BAG angekündigten Konzept zur Publikation von «schweregradbereinigten Fallkosten» im Rahmen von Artikel 49 Absatz 8 KVG gegenwärtig genügend Bestrebungen im Gange [6].

Drittens: Weil die Kantone bei der Spitalplanung sicherstellen müssen, dass die Leistungsaufträge an Spitäler vergeben werden, welche den WZW-Anforderungen genügen, wird das Spital Benchmarking auch für die Wirtschaftlichkeitsprüfung im Rahmen der Spitalplanung verwendet. Dieser zweite Verwendungszweck des Spital Benchmarkings wird in dieser Studie nicht weiter behandelt.

1.3 Methodik

Die Kapitel 2.1 bis 2.3 basieren auf einer Literaturanalyse, welche auf drei Themenbereiche fokussierte:

1. Die Verwendung von Gewichten in der angewandten Statistik
2. Die Berechnung der SwissDRG Kostengewichte und Fallnormkosten
3. Die Vergleichbarkeit der Fallnormkosten im SwissDRG Tarifsystem zwischen den Spitälern

Die Literaturanalyse erfolgte strukturiert anhand von Schlagwörtern in wissenschaftlichen Journals, Google Scholar, Fachpublikationen, Lehrbüchern und Homepages (z.B. SwissDRG).

Ausgehend von den drei untersuchten Themenbereichen erfolgte die Literaturanalyse in zwei Schritten. Im ersten Schritt haben wir die Fachliteratur aus verschiedenen Gebieten der angewandten Statistik (empirische Sozialforschung, Ökonometrie, etc.) hinsichtlich der möglichen Gewichtungsmotive ausgewertet und die aus der jeweiligen Gewichtung resultierenden Vor- und Nachteile erarbeitet. Im zweiten Schritt haben wir die Literatur zur Berechnung der SwissDRG Kostengewichte und der Fallnormkosten analysiert und die empirisch identifizierten Mängel bzw. Grenzen von SwissDRG hinsichtlich der Vergleichbarkeit der Fallnormkosten zwischen den Spitälern aufgearbeitet, in einer Übersicht zusammengestellt und kritisch gewürdigt.

Die Beurteilung der Gewichtung im Rahmen des Spital Benchmarkings im Kapitel 2.4 erfolgte anhand der Synthese der Resultate aus der Literaturanalyse und eigenen konzeptionellen Überlegungen. Zur Verdeutlichung unserer Schlussfolgerungen wurde ein fiktives Benchmarking-Beispiel für 10 Spitäler entwickelt.

2 Vor- und Nachteile einer Gewichtung im Spital Benchmarking

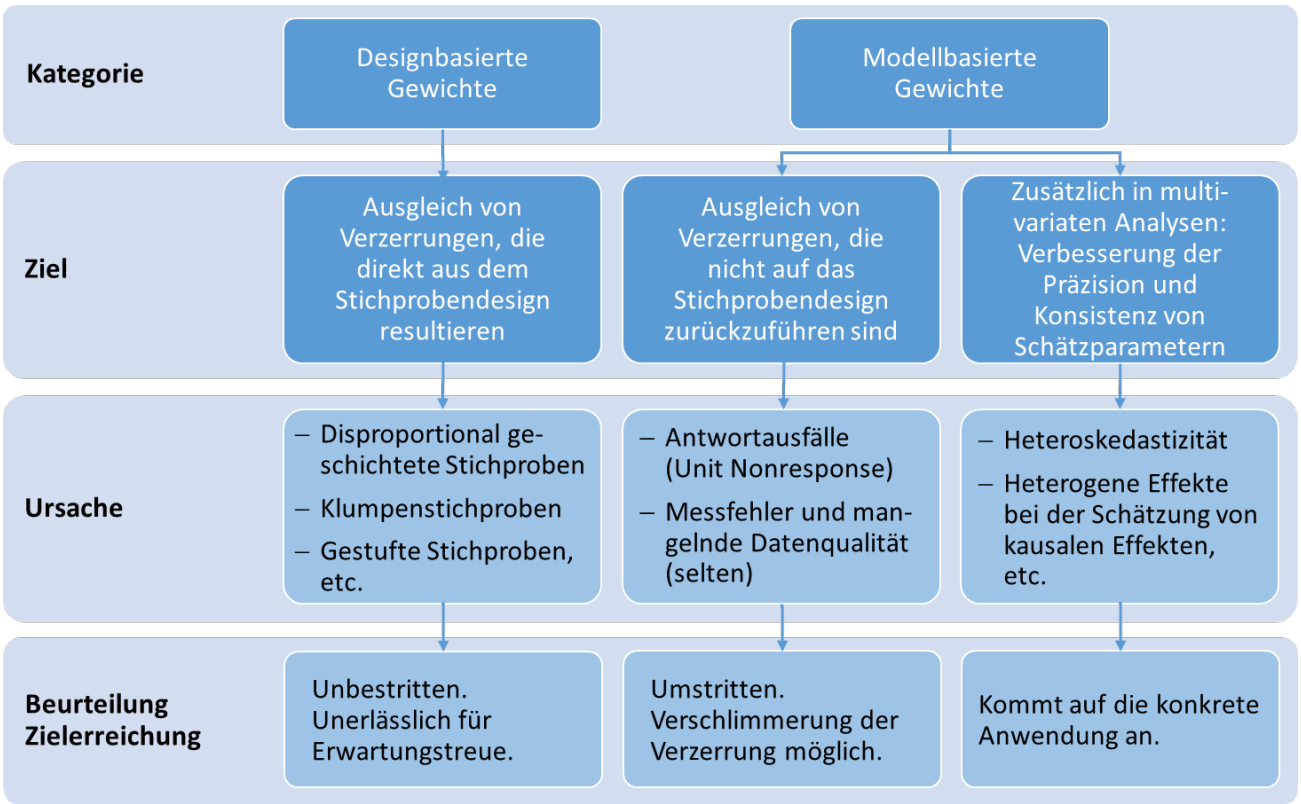
2.1 Gewichtung in der angewandten Statistik

Im Allgemeinen versteht man unter Gewichtung in der angewandten Statistik den Versuch, die Repräsentativität einer Stichprobenerhebung zu erhöhen, um letztendlich anhand von Stichprobenkennwerten, wie zum Beispiel Mittelwerten, Aussagen über die Grundgesamtheit treffen zu können. Primäres Ziel einer Gewichtung ist es somit möglichen Verzerrungen, welche direkt aus dem Stichprobendesign oder dem Antwortverhalten befragter Einheiten resultieren, entgegenzuwirken.

Im technischen Sinne bedeutet die Gewichtung, dass jede Beobachtungseinheit mit einem multiplikativen Faktor versehen wird. Im Datensatz entsteht eine neue (Gewichtungs-)Variable, welche jeder beobachteten Einheit ihr Gewicht zuordnet. Die Gewichte sind dabei beliebig skalierbar und lassen sich auch als ganze positive Zahlen ausdrücken, die angeben, wie häufig die einzelne Beobachtung zur Berechnung des Kennwertes berücksichtigt bzw. im Datensatz «virtuell repliziert» wird. Dadurch gehen Messwerte einzelner Beobachtungseinheiten mit veränderter relativer Bedeutung in die Berechnung der Stichprobenkennwerte ein.

In der angewandten Statistik existieren zahlreiche Gewichtungsverfahren. Über den Sinn und Zweck dieser Verfahren wird jedoch schon lange kontrovers diskutiert [7]. Aktuell verwendete Gewichtungsverfahren lassen sich in zwei Kategorien unterteilen: designbasierte und modellbasierten Verfahren [8]. Im Folgenden werden diese beiden Gewichtungskategorien vorgestellt und die Vor- und Nachteile, welche aus design- bzw. modellbasierten Gewichtungsverfahren resultieren können, diskutiert (vgl. Abbildung 1).

Abbildung 1: Übersicht Gewichtungsverfahren, Zweck und Würdigung



2.1.1 Zwei Kategorien von Gewichtungsverfahren

Designbasierte Gewichtung

Designbasierte Gewichte dienen dem Ausgleich von Verzerrungen, die sich direkt aus dem Design der Stichprobenauswahl ergeben. Häufig wird in der Praxis aus Praktikabilitäts- und Kostengründen von der Ziehung einer einfachen Zufallsstichprobe (random sample) abgesehen und ein komplexeres Stichprobendesign verwendet. Anders als bei der einfachen Zufallsstichprobe, unterscheiden sich bei komplexen Stichprobendesigns die Auswahlwahrscheinlichkeiten der Beobachtungseinheiten voneinander, sodass die Stichprobe kein unverzerrtes Abbild der Grundgesamtheit mehr liefert. Dies hat zur Folge, dass Schätzer für Populationsparameter in der Regel nicht mehr erwartungstreu, sondern verzerrt sind, d.h. einen systematischen Fehler aufweisen. Daher sind Korrekturmassnahmen notwendig, wenn von den geschätzten Stichprobenkennwerten auf die Populationsparameter der Grundgesamtheit geschlossen werden soll.

Komplexere Stichprobendesigns, welche in der Praxis von zentraler Bedeutung sind, sind zum Beispiel disproportional geschichtete Stichproben, Klumpenstichproben oder gestufte Stichproben. Grundsätzlich ist diesen Designs gemein, dass die Auswahl der Untersuchungseinheiten in einem mehrstufigen Verfahren erfolgt. Das heisst, anders als bei der einfachen Zufallsstichprobe, wird nicht direkt aus der Grundgesamtheit gezogen. So werden zum Beispiel bei einer disproportional geschichteten Stichprobe basierend auf Merkmalen wie Geschlecht, Alter, Nationalität oder Wohnkanton zunächst Gruppen gebildet, aus denen dann im Anschluss einfache Zufallsstichproben gezogen werden. Da die Stichprobengrösse völlig losgelöst von der jeweiligen Gruppengrösse in der Grundgesamtheit bestimmt werden kann, können so Individuen aus kleinen Bevölkerungsgruppen überproportional häufig Eingang in eine Stichprobe finden. Dies ermöglicht trotz beschränkter Gruppengrösse, repräsentative Auswertungen für kleine Bevölkerungsgruppen, führt aber zu systematischen Verzerrungen bei einer ungewichteten Auswertung auf Gesamtstichprobenebene. Ein Beispiel dafür ist die Schweizerische Gesundheitsbefragung, bei der Personen aus kleinen Kantonen gezielt «zu häufig» befragt («oversampled») werden, um repräsentative Auswertungen auf kantonaler Ebene zu ermöglichen [9].

Designbasierte Gewichte zielen darauf ab, Verzerrungen, die aus solchen komplexen Stichprobendesigns mit mehrstufigem Auswahlverfahren resultieren, auszugleichen. Sie werden als Faktor proportional zur inversen Auswahlwahrscheinlichkeit einer Beobachtungseinheit gebildet. Damit fliessen überrepräsentierte Beobachtungseinheiten weniger stark und unterrepräsentierte Beobachtungseinheiten stärker in die Berechnung der Kennwerte ein [10]. Rothe und Wiedenbeck zeigen, dass Mittelwertschätzer mit Designgewichten theoretisch erwartungstreu sind, sofern kein Noncoverage vorliegt, d.h. sofern alle Beobachtungseinheiten eine Auswahlwahrscheinlichkeit von grösser Null haben, *und* sich die Auswahlwahrscheinlichkeiten korrekt aus dem Stichprobendesign herleiten lassen [7].

Modellbasierte Gewichtung

Zu den modellbasierten Gewichtungsverfahren zählen alle Methoden, die nicht durch das Stichprobendesign legitimiert werden. Wir haben sie entsprechend ihrem Anwendungsgebiet (deskriptiv / multivariat) und den damit verbundenen Zielen in zwei Gruppen unterteilt. Das ursprüngliche Ziel von modellbasierten Gewichtungsverfahren liegt im Ausgleich von Verzerrungen, welche nachgelagert an die Stichprobenziehung entstehen, etwa durch Antwortausfälle (Nonresponse) oder seltener auch Messfehler.

Bei den Antwortausfällen unterscheidet man zwischen Unit Nonresponse und Item Nonresponse. Unit Nonresponse liegt vor, wenn Personen aus der Stichprobe nicht befragt werden konnten, weil sie nicht erreichbar

waren, nicht teilnehmen konnten (z.B. durch Krankheit oder Sprachbarrieren) oder die Teilnahme verweigerten. Da in diesem Fall zu keiner Frage eine Antwort vorliegt, handelt es sich um einen vollständigen Antwortausfall. Dem gegenüber steht der partielle Antwortausfall bzw. die Item Nonresponse. Diese liegt vor, wenn die Befragten auf einzelne oder mehrere Fragen keine Antworten gaben, weil sie sensible Informationen hätten preisgegeben müssen oder Fragen schlichtweg überlesen wurden. Da in diesem Fall gewisse Informationen zu den Befragten vorhanden sind, die zur Ergänzung fehlender Informationen genutzt werden können, wird bei Item Nonresponse in der Regel ein Ergänzungsverfahren (Imputation) gegenüber einem Gewichtungsverfahren vorgezogen [10]. Dies gilt nicht zuletzt auch aus Effizienzgründen [11]. Das Verwerfen ganzer Fälle wird im Allgemeinen sehr kritisch gesehen [10]. Aus diesen Gründen werden wir uns im Folgenden auf den Fall des Unit Nonresponse beschränken.

Unit Nonresponse ist problematisch, weil sie im Allgemeinen nicht zufällig ist, sondern von bestimmten Personeneigenschaften abhängt. In Folge unterscheiden sich Teilnehmer von Nichtteilnehmern und die Stichprobenkennzahlen sind in der Regel verzerrt [10]. Sogenannte Nonresponse-Gewichte zielen darauf ab, diese Verzerrungen auszugleichen, indem jede Beobachtungseinheit mit ihrer inversen Teilnahmewahrscheinlichkeit gewichtet wird. So werden Personen(gruppen) mit tiefer Teilnahmewahrscheinlichkeit höher gewichtet als Personen mit hoher Teilnahmewahrscheinlichkeit. Theoretisch sind dann unverzerrte Stichprobenkennwerte schätzbar. Das Problem in der Praxis besteht jedoch darin, dass die Antwortwahrscheinlichkeiten nicht beobachtet sind. Sie müssen aus der *Stichprobe geschätzt* werden.

Einen guten Überblick über *verschiedene Schätztechniken* zur Schätzung der Antwortwahrscheinlichkeiten findet sich in Brick (2013) [12]. Im Wesentlichen kann zwischen zwei Arten von Schätztechniken unterschieden werden: (1) das Bilden von Gruppen, innerhalb welcher für die Beobachtungseinheiten gleiche Antwortwahrscheinlichkeiten erwartet werden, und (2) das Schätzen eines expliziten Modells für die individuellen Antwortwahrscheinlichkeiten (z.B. Logit). Allen Techniken ist gemein, dass für eine korrekte Schätzung der Antwortwahrscheinlichkeiten alle Variablen bekannt sein müssen, die die Antwortwahrscheinlichkeiten beeinflussen (missing at random assumption). Das gilt auch für diejenigen Beobachtungen, die nicht an der Befragung teilgenommen haben. Diese starke Annahme ist in der Praxis im besten Fall näherungsweise erfüllt. Meist liegen für die Nichtteilnehmer nur Basisinformationen aus Registern vor [10]. Besser sieht es bei Panelbefragungen aus. Hier können alle Informationen aus der Befragung vom Vorjahr für die Schätzung verwendet werden (z.B. [13], [14]). Überprüfen kann man die Güte der geschätzten Antwortwahrscheinlichkeiten jedoch nicht.

Eine Alternative zur Schätzung der Antwortwahrscheinlichkeiten ist die sogenannte *Kalibrierung,* auch *Redressment* oder *Poststratifikation* genannt (manchmal wird sie auch zusätzlich zu Design- und Nonresponse-Gewichten angewendet). Unter *Kalibrierung* versteht man die Anpassung von Merkmalen in der Stichprobe an Eckdaten der amtlichen Statistik. Erreicht wird diese Anpassung mittels sogenannten Redressment Gewichten, die oft auch mehrere Merkmale miteinander verknüpfen, wie z.B. Gemeinde x Altersgruppe x Geschlecht. Die Wahl der Merkmale (sog. Redressment Merkmale) ist oft willkürlich und wird durch die in öffentlichen Statistiken vorhandenen Merkmale begrenzt. Die zugrundeliegende Idee ist, dass durch Anpassung der Redressment Merkmale an Populationswerte auch die sogenannten «passiven Merkmale» – so werden die Merkmale bezeichnet, die nicht zur Kalibrierung verwendet wurden – näher an jene der Gesamtpopulation herankommen. Dies funktioniert aber nur dann, wenn die Antwortwahrscheinlichkeit ausschliesslich von den Redressment Merkmalen abhängt, eine Annahme, die nicht überprüft werden kann.

Die Anwendung von Gewichten mit dem Ziel, Verzerrungen durch **Messfehler** auszugleichen, ist kaum verbreitet. Eine der wenigen Studien dazu ist jene von Durrand und Skinner (2006) [11]. Die Autoren zeigen, dass mittels Propensity Score Weighting für Messfehler korrigiert werden kann, sofern für eine Teilstichprobe die

korrekten Zahlen bekannt sind. Dies verdeutlichten sie am Beispiel von aus Umfragen erhobenen Stundenlöhnen. Dazu schätzten sie zuerst sogenannte Propensity Scores. Die Propensity Scores gaben an, wie ähnlich sich die Personen in den beobachteten Eigenschaften sind und ermöglichten die Identifikation von sogenannten «nächsten Nachbarn». So wurde für jede Person mit fehlenden korrekten Lohninformationen der «nächste Nachbar» unter den Personen mit korrekten Lohninformationen bestimmt, was schliesslich die Berechnung von Gewichten ermöglichte. Je öfter eine Person als «nächster Nachbar» identifiziert wurde, desto höher ist ihr Gewicht in der Schätzung des Durchschnittslohnes. Eine weitere Anwendung findet sich im Rahmen von Expertenbefragungen (z.B. mittels Delphi-Methode) zum Ausgleich von mangelnder Datenqualität. Loveridge et al. (1995) schlagen in diesem Zusammenhang vor, die Antworten von kompetenteren Experten stärker zu gewichten, als jene von weniger kompetenten Experten. Dabei gehen sie implizit von der Annahme aus, dass kompetentere Experten bessere Schätzungen abgeben als andere ([15] in [10]). Ob und wie Gewichte im Fall eines Messfehlers verwendet werden können, hängt jedoch von der konkreten Anwendung ab.

In der **multivariaten Statistik** kommen zusätzliche Gewichtungsmotive hinzu. Zum Beispiel kann bei OLS Regressionen die Heteroskedastizität in den Residuen mit Gewichten reduziert werden, wodurch die Präzision der Schätzparameter erhöht wird (Homoskedastizität ist eine Voraussetzung für unverzerrte Standardfehler) [16]. Weitere Anwendungen finden sich im Zusammenhang mit der Schätzung von kausalen Effekten (Ursache-Wirkungs-Zusammenhängen) [17]. Grundsätzlich ist es jedoch so, dass Gewichte in multivariaten Modellen redundant werden, sobald alle Variablen, welche die Auswahl- und Antwortwahrscheinlichkeiten beeinflussen, als Kontrollvariablen mit ins Modell einfliessen. Da eine Gewichtung in diesem Fall lediglich zu grösseren Standardfehlern führt, wird eine ungewichtete einer gewichteten Schätzung vorgezogen. Eine umfassende Diskussion bezüglich der Pros und Contras von Gewichtung in der multivariaten Statistik findet sich zum Beispiel in Gelman (2007) [18]. Weiter werden wir an dieser Stelle nicht auf die Anwendung von Gewichten in der multivariaten Statistik eingehen, da dieser Anwendungsbereich für das Spital Benchmarking, welches auf rein deskriptiven Grössen basiert, nicht relevant ist.

2.1.2 Kritische Würdigung

Die Anwendung von *designbasierten Gewichten* für erwartungstreue Schätzer ist in der Theorie unbestritten und in der Praxis, insbesondere in deskriptiven Anwendungen, weit verbreitet. Dennoch gilt es für praktische Anwendungen zu bedenken, dass Schätzer nur dann unverzerrt bzw. erwartungstreu sind, wenn die Auswahlwahrscheinlichkeiten für alle Beobachtungen korrekt aus dem Stichprobendesign hergeleitet werden können. Gerade bei komplexen Stichprobendesigns ist dies jedoch meist nur näherungsweise möglich [7], [16]. In der empirischen Forschung scheint man sich aber einig, dass designbasierte Gewichte in diesem Fall eher Verzerrung korrigieren, als dass sie zusätzliche Verzerrungen verursachen.

Modellbasierte Gewichte werden zwar in den meisten populationsrepräsentativen Surveys berechnet und ausgewiesen, aber ihre Anwendung in der empirischen Forschung ist hoch umstritten. Denn analog zu den designbasierten Gewichten, welche die Kenntnis der korrekten Auswahlwahrscheinlichkeiten voraussetzen, setzt die Korrektur von Verzerrungen durch Nonresponse oder Messfehler die Kenntnis der korrekten Antwortwahrscheinlichkeiten bzw. die Höhe des Messfehlers voraus. Diese sind jedoch gänzlich oder zumindest teilweise unbeobachtet, d.h. sie können nicht aus dem Stichprobendesign hergeleitet werden, sondern müssen aus den Daten geschätzt werden. Für eine korrekte Schätzung der Antwortwahrscheinlichkeiten bzw. des Messfehlers müssen alle Variablen bekannt sein, die die Antwortwahrscheinlichkeiten bzw. den Messfehler beeinflussen, eine starke Annahme, die zumeist nicht überprüft werden kann.

In der Praxis hofft man oft darauf, dass modellbasierte Gewichte etwaige Verzerrungen zumindest verringern [10], [12]. Tatsächlich gibt es aber verschiedene Studien, die zeigen, dass die Verwendung von «falschen» Gewichten (Gewichte, die diese Voraussetzung nicht erfüllen) die Verzerrungen auch vergrössern können. Zum Beispiel wurde die Validität von Redressment Gewichten getestet, indem Variablen, welche selbst keine Redressment Merkmale darstellen, aber aus öffentlichen Statistiken bekannt sind, mit Redressment Gewichten «korrigiert» und dann mit den entsprechenden Werten aus öffentlichen Statistiken verglichen wurden. Diese Studien kommen zum Schluss, dass nicht vorhergesagt werden kann, ob Redressment Gewichte deskriptive Statistiken «verbessern» oder sie weiter «verschlechtern» [7].

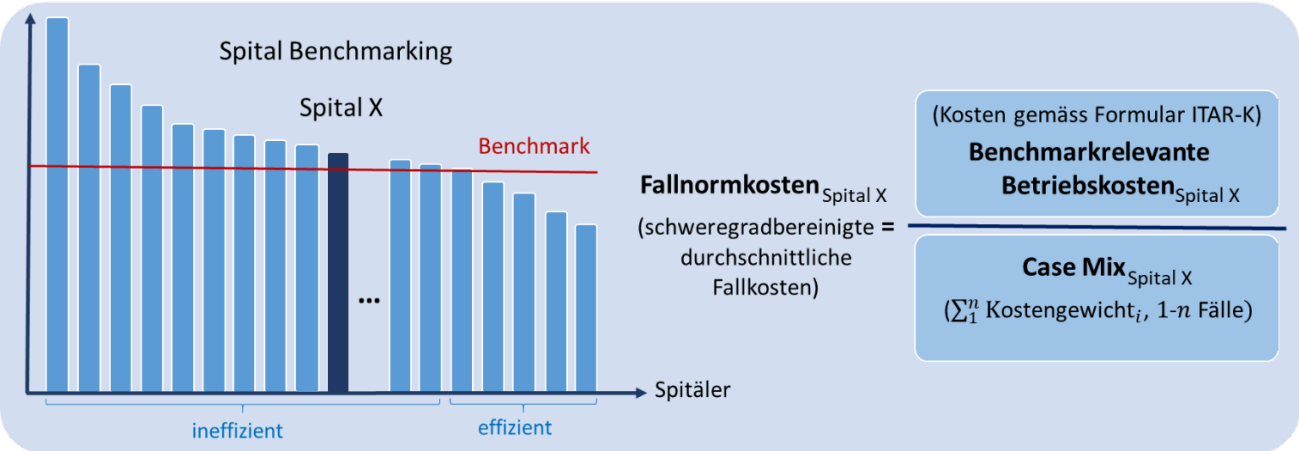
In der multivariaten Statistik hat sich deshalb als Best Practice etabliert, dass neben gewichteten Resultaten immer auch die ungewichteten Resultate zum Vergleich ausgewiesen werden. Bestehen grosse Unterschiede, sollte die Ursache genauer untersucht und erklärt werden [16].

2.2 Spital Benchmarking: Zweck und Herausforderungen

Das Ziel des Spital Benchmarkings ist die Ermittlung eines effizienten, leistungsgerechten Preises, wie er sich in einem freien Markt unter idealen Bedingungen durch echten Preiswettbewerb einstellen würde. In der Rechtsprechung und der Praxis wird dazu ein Benchmarking angewendet, das sich auf «schweizweiten Betriebsvergleichen zu Kosten» (Art. 49 Abs. 8 KVG) abstützt.

Basierend auf der Annahme, dass die Tarifstruktur SwissDRG in der Lage ist, die Leistungen aller Spitäler vergleichbar zu machen, basiert das Spital Benchmarking konzeptionell auf den sogenannten Fallnormkosten. Diese ergeben sich aus den benchmarkrelevanten Betriebskosten geteilt durch den Case Mix. Letzterer entspricht der Summe aller Kostengewichte innerhalb eines Spitals (Vgl. Abbildung 2). Die Fallnormkosten sind somit als durchschnittliche schweregradbereinigte Fallkosten von stationären Behandlungen innerhalb eines Spitals zu verstehen. Unterschiede in den Fallnormkosten zwischen den Spitälern reflektieren unter der obigen Annahme einzig und allein Unterschiede in der Effizienz der Leistungserbringung. Mit der Wahl des Perzentils (Benchmark) kann somit der Basispreis beim erwünschten Grad an Effizienz festgelegt werden.

Abbildung 2: Fallnormkosten im Spital Benchmarking



Wie verschiedene wissenschaftliche Gutachten heute jedoch zeigen, kann die SwissDRG Tarifstruktur die Vergleichbarkeit zwischen den Leistungen der Spitäler nicht wie erhofft herstellen. Folglich stellen die Fallnormkosten keine geeignete Grundlage für die Effizienzvergleiche im Rahmen des Spital Benchmarking dar, obgleich

die aktuelle Rechtsprechung genau diesen Vergleich explizit vorschreibt. Eine Unterteilung in effiziente und ineffiziente Spitäler ist nicht möglich, weil Unterschiede zwischen Spitälern nur teilweise auf Effizienzunterschiede zurückzuführen sind. Weitere Unterschiede ergeben sich zum Beispiel durch leistungs- und patientenbezogene Unterschiede zwischen den Spitälern (z.B. [19]).

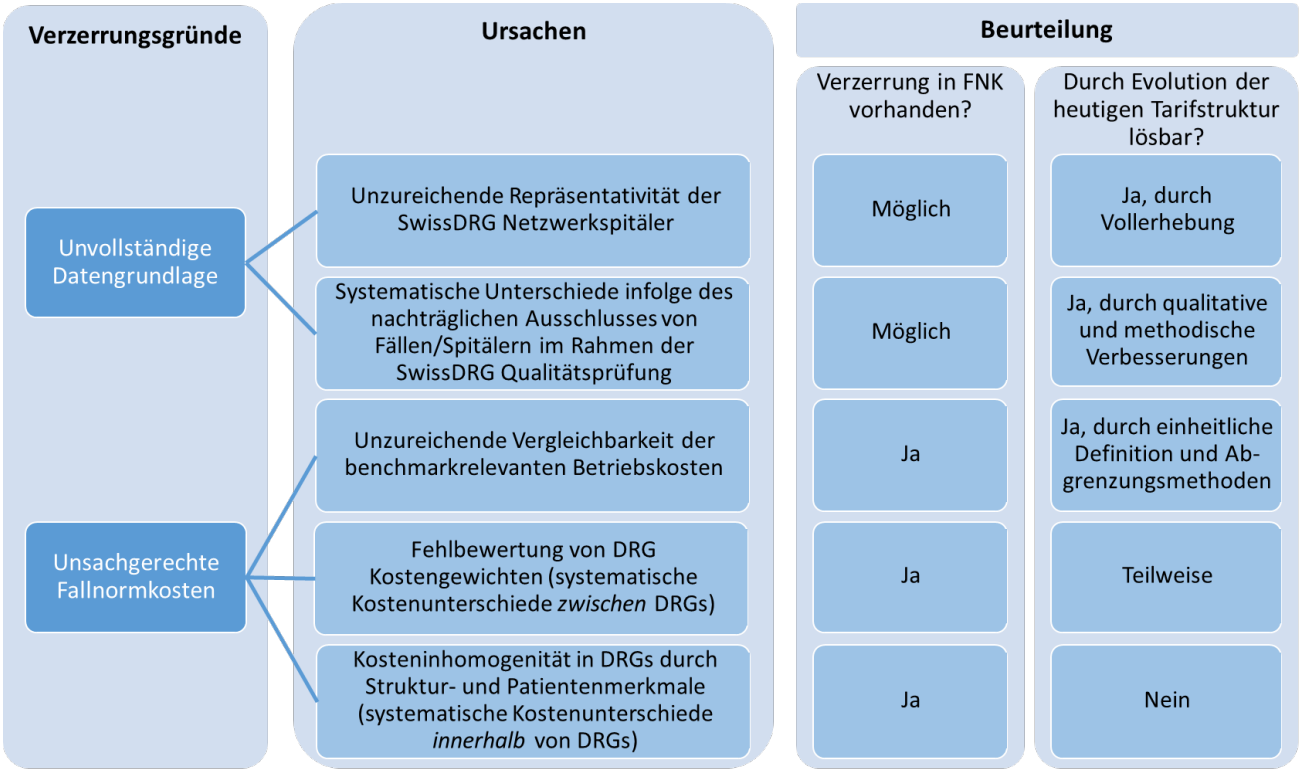
Vor diesem Hintergrund wird aktuell darüber diskutiert, ob differenzierte Baserates notwendig sind oder ob alternativ eine Gewichtung der Spitäler vorgenommen werden kann, welche die Verzerrungen in den Fallnormkosten auszugleichen vermag.

Im Folgenden werden zunächst die in den wissenschaftlichen Gutachten identifizierten Gründe für mögliche Verzerrungen in den Fallnormkosten zusammengetragen und diskutiert.

2.3 Validität der Fallnormkosten als Grundlage für das Spital Benchmarking

In der Literatur wurden fünf mögliche Ursachen für eine Verzerrung in den Fallnormkosten identifiziert, welche sich zwei übergeordneten Themenbereichen zuordnen lassen (s. Abbildung 3).

Abbildung 3: Ursachen für eine Verzerrung der SwissDRG Fallnormkosten



Unvollständige Datengrundlage

Eine unvollständige Datengrundlage resultiert vor allem aus der Tatsache, dass aktuell bei der Berechnung der DRG Kostengewichte immer noch auf eine Stichprobe von Netzwerkspitälern zurückgegriffen wird und nachträglich im Rahmen der Qualitätsprüfung von SwissDRG weitere Spitäler und Fälle ausgeschlossen werden [20]. Anders als bei einer Vollerhebung und anschliessender Verwendung des kompletten Datensatzes kann es deshalb bei der Kalkulation der Kostengewichte zu systematischen Verzerrungen kommen. Dies ist insofern

problematisch, als dass die Kostengewichte über die Normierung der Fallkosten in den Spitalvergleich mit eingehen.

Unzureichende Repräsentativität der SwissDRG Netzwerkspitäler

Von einer unzureichenden Repräsentativität der Netzwerkspitäler wird gesprochen, wenn sich die DRG Netzwerkspitäler (d.h. Spitäler, die Daten an SwissDRG liefern) systematisch von den Spitälern, die keine Daten an SwissDRG liefern, unterscheiden. Dies führt dazu, dass die Stichprobe der Netzwerkspitäler bezüglich des gesamten Spektrums der akutsomatischen Leistungen und der damit verbundenen Fallkosten nicht mehr repräsentativ ist. Rückschlüsse von der Stichprobe auf die Grundgesamtheit der für SwissDRG relevanten Akutspitäler sind dann nicht mehr ohne weiteres möglich.

Obgleich es sich hierbei um einen möglichen Verzerrungsgrund handelt, hat dieser in der Praxis praktisch kaum noch Relevanz. Mit der Einführung der Datenlieferungspflicht für das Datenjahr 2012 [21] ist die Anzahl der Netzwerkspitäler sprunghaft angestiegen. Während im Datenjahr 2011 nur 68 der 204 nach SwissDRG abrechnenden Spitäler Daten an SwissDRG lieferten, waren es im Datenjahr 2012 bereits 102 von 197 Spitälern [22]. Auch in den Folgejahren konnte die Datenbasis immer weiter ausgebaut werden. Im Datenjahr 2017 umfasste die Stichprobe der Netzwerkspitäler 117 der 177 für SwissDRG relevanten akutsomatischen Spitäler [20]. Mit einem Abdeckungsgrad von 66.1% war man damit zwar noch weit von einer Vollerhebung entfernt, aber diese Zahl relativiert sich deutlich, sobald berücksichtigt wird, dass die Datenlieferung der Netzwerkspitäler bereits knapp 90% der für SwissDRG relevanten akutsomatischen Fälle umfasste [20]. Langfristig ist eine Vollerhebung angestrebt [21]. Derzeitige Bestrebungen des Bundesrates, finanzielle Strafen für Spitäler einzuführen, die «ihre Kosten- und Leistungsdaten verspätet, in unzureichender Qualität oder gar nicht für die Ermittlung des Benchmark Wertes […] bereitstellen» [23], könnten dazu führen, dass dieses Ziel auch in naher Zukunft erreicht wird.

Inwiefern sich im aktuellen Datenjahr Netzwerkspitäler von den nicht an SwissDRG liefernden Spitälern unterscheiden, lässt sich auf Basis vorliegender Informationen nicht beurteilen. Ein Abgleich von den in der Krankenhausstatistik enthaltenen Spitalmerkmalen, könnte jedoch erste Hinweise darüber liefern, ob sich liefernde und nicht liefernde Spitäler systematisch unterscheiden. Weitere nicht in der Krankenhausstatistik enthaltene Spitalmerkmale wären künftig zusätzlich durch die SwissDRG AG zu prüfen.

Nachträglicher Ausschluss von Fällen und Spitälern im Rahmen der Qualitätsprüfung

Im Rahmen der Qualitätsprüfung von SwissDRG werden weitere Fälle, aber auch ganze Spitäler, von der Datenbasis ausgeschlossen (z.B. [24], [25], [20]). Im Datenjahr 2017 betraf dies 147'386 der 1‘275‘168 für SwissDRG relevanten Fälle, sodass nunmehr nur noch 77.8% der relevanten akutsomatischen Fälle bei der Berechnung der DRG Kostengewichte Berücksichtigung fanden [20]. 11 Spitäler wurden aus dem Datensatz ganz gelöscht [20]. Als Gründe für diesen Ausschluss wurden die Abgrenzungsproblematik OKP / VVG (private oder halbprivate Fälle), Item Nonresponse (Fälle ohne Vollkosten) und mögliche Datenfehler (unplausible Fälle) genannt [20]. Gerade der letzte Grund stellt mit rund 121'858 ausgeschlossenen Fällen möglichweise das grösste Problem dar, nicht zuletzt auch, weil ganze Spitäler ausgeschlossen wurden. Besteht ein Zusammenhang zwischen Datenqualität und Fallkosten, was zum Beispiel durch eine Häufung von Codier-Fehlern bei sehr komplexen Fällen nicht unwahrscheinlich ist, dann kommt es zu einem systematischen Ausschluss von Fällen. In Folge werden Kostengewichte (inkl. der Zu- und Abschläge) und Fallnormkosten nicht mehr korrekt berechnet, was die Vergleichbarkeit der Spitäler untereinander beeinträchtig, vor allem da sich Spitäler hinsichtlich ihrer Fallstruktur unterscheiden [26]. Der Ausschluss ganzer Spitäler kann zu zusätzlichen systemati-

schen Verzerrungen führen, zum Beispiel, wenn sich die anrechenbaren Betriebskosten infolge unterschiedlicher Leistungsaufträge unterscheiden oder Manipulationsanreize, wie die bewusste Lieferung einer «schlechten» Datenqualität, durch Spitäler entsprechend genutzt werden. Auch die anderen beiden von SwissDRG angeführten Ausschlussgründe (Abgrenzungsproblematik OKP / VVG, Item Nonresponse) können zu einer Verzerrung führen, wenn die resultierende Datenbasis nicht mehr repräsentativ für die für SwissDRG relevanten akutsomatischen Fälle und Spitäler ist. Aufgrund der geringeren Fallzahlen stufen wir hier die erwartete Verzerrung jedoch als deutlich geringer, wenn nicht sogar vernachlässigbar klein, ein.

Wissenschaftliche Studien, die das Ausmass der Verzerrung durch den nachträglichen Ausschluss unplausibler Fälle bzw. ganzer Spitäler abschätzen, sind uns nicht bekannt. Daher lässt sich an dieser Stelle nur festhalten, dass Verzerrungen aus theoretischer Sicht möglich sind, aber die Repräsentativität des letztendlich verwendeten Datensatzes noch zu beurteilen ist. Mehr Transparenz seitens der SwissDRG AG wäre dafür erforderlich.

Langfristig betrachtet schätzen wir das Ausmass der Verzerrung durch den nachträglichen Ausschluss von Spitälern und Fällen als wenig relevant ein. Grund dafür sind einerseits die zu erwartenden Verbesserung der Datenqualität. Über die Zeit findet ein Lern- und Gewöhnungsprozess statt. Daher sollten Datenfehler künftig seltener auftreten, sodass insgesamt mit weniger unplausiblen Fällen zu rechnen ist. Auch das Item Nonresponse Problem (Fälle ohne Vollkosten) wird sich möglichweise reduzieren. Eine Intensivierung der bereits erfolgenden Praktiken von SwissDRG, um Unstimmigkeiten in den Daten in Absprache mit den Spitälern aufzuklären, könnte darüber hinaus zu einer Verbesserung der Datenqualität beitragen, ebenso wie die bereits erwähnten, geplanten Strafzahlungen des Bundesrates, welche von der gelieferten Datenqualität abhängen sollen. Andererseits können möglicherweise auch methodische Verbesserungen zu einem Ausbau der Datenbasis führen. So ist zum Beispiel von der SwissDRG AG zu prüfen, ob mögliche Lücken im Datensatz durch Imputationsmethoden ergänzt werden können. Mit zunehmend mehr Informationen, die allenfalls auch aus den Vorjahren zur Verfügung stehen, führt dies wahrscheinlich zu besseren Ergebnissen, als wenn komplette Fälle ignoriert werden [10], wie dies aktuell der Fall ist.

Unsachgerechte Fallnormkosten

Unsachgerechte Fallnormkosten werden in der Literatur vor allem auf zwei Ursachen zurückgeführt: die Fehlbewertung von DRG Kostengewichten und die Kosteninhomogenität innerhalb von DRGs (vgl. [4], [5], [19], [27]–[29]). Streng genommen führt jedoch noch eine dritte Ursache zu unsachgerechten Fallnormkosten – die inkorrekte Berechnung der benchmarkrelevanten Betriebskosten [19], [30]. Der Grund ist, dass diese sowohl in den Zähler (direkt) als auch den Nenner (indirekt über Kostengewichte) der Fallnormkosten eingehen.

Unzureichende Vergleichbarkeit der benchmarkrelevanten Betriebskosten

Die Vergleichbarkeit der benchmarkrelevanten Betriebskosten setzt voraus, dass die Kosten für stationäre Leistungen gemäss KVG, inklusive der Anlagenutzungskosten (ANK), korrekt und einheitlich berechnet werden. Zur Bewertung der ANK hat der Bundesrat jedoch Vorgaben erlassen, die in ihrer Herleitung dem KVG widersprechen und zu einer «durch effizienzfremde Faktoren und Zufälligkeiten» verzerrten Bewertung der ANK führen [31]. Abgrenzungsprobleme ergeben sich insbesondere bei den Kosten für gemeinwirtschaftliche Leistungen, da bis heute keine einheitliche Definition dieser Leistungen vorliegt [19]. Weitere Abgrenzungsprobleme liegen bei den Kosten aus liegeklassenbedingten Mehrleistungen, Arzthonoraren bei Zusatzversicherten sowie den Kosten für Forschung und Lehre [19]. Mangels einheitlicher Abgrenzungsmethoden wird hier auf Normabzüge zurückgegriffen [1], [32], was zu weiteren Unschärfen führt.

Da mit dem vom BAG für 2020 angekündigten Konzept zur Publikation von «schweregradbereinigten Fallkosten» bereits auf die unsachgerechte Abgrenzung und mangelnde Vergleichbarkeit der Betriebskosten reagiert

wird [6], sehen wir an dieser Stelle von einer weiteren Diskussion dieser Verzerrungsursache ab. Es ist offensichtlich, dass durch einheitliche Definitionen und Abgrenzungsmethoden zeitnah eine grundlegende Verbesserung erzielt werden kann und dieser Verzerrungsgrund möglicherweise entfällt.

Fehlbewertung von DRG Kostengewichten

Eine korrekte Bewertung verlangt, dass das einer Fallgruppe zugeordnete Kostengewicht die erwartete durchschnittliche Fallschwere dieser Gruppe abbildet. Bestehen Fehlbewertungen und unterscheiden sich diese zwischen den einzelnen DRGs, dann sind einige DRGs lukrativ und andere nicht. Je nach Leistungsangebot eines Spitals können sich so systematische Gewinne oder Verluste ergeben.

In diesem Zusammenhang wird in der Schweiz insbesondere auf den Kompressionseffekt verwiesen (z.B.[4], [19], [27]). Dieser bezeichnet in DRG-Systemen die tendenzielle Überbewertung von einfachen (Standard-) Fällen bei gleichzeitiger Unterbewertung von aufwendigen (Komplex-) Fällen. Als Gründe für den Kompressionseffekt in der Schweiz werden vor allem tiefe Fallzahlen (bezogen auf eine Fallgruppe) und Outlier-Korrekturen benannt [28], [33]. Outlier werden von SwissDRG bei der Berechnung der Kostengewichte nicht berücksichtigt, sodass die medizinischen Kosten der Outlier, welche nicht über Zu- und Abschläge vergütet werden, im Mittel gerade denjenigen der Normallieger entsprechen müssen. Dies ist in der Regel nicht der Fall, weshalb es je nach Verteilung der Kurz- und Langlieger zu systematischen Gewinnen und Verlusten von DRGs kommen kann. Aber auch Mängel bei der Leistungs- und Kostenerfassung sowie Unschärfen bei der Zuordnung der Fälle zu gewissen Fallgruppen stellen mögliche Ursachen dar (vgl. [34], [19]).

Empirisch wurde der Kompressionseffekt mehrfach belegt (z.B. [19], [35]). Zudem wurden weitere Fehlbewertungen von DRG Kostengewichten nachgewiesen. Diese führten zu einer systematischen Benachteiligung von Universitäts- und Kinderspitälern [28], [29]. So waren zum Beispiel gemäss Fallkostenstatistik des BFS 2012 für Universitätsspitäler rund ein Fünftel der unberücksichtigten Kostenunterschiede auf systematische Unterschiede zwischen den DRGs zurückzuführen, was hauptsächlich durch Langlieger bedingt war [28]. Bei Kinderspitälern wurden gemäss Fallkostenstatistik des BFS 2014 ebenfalls deutliche Mehrkosten aufgrund des speziellen Leistungsangebots (rund 100 CHF höhere Fallnormkosten als beim Schweizer Durchschnitt) nachgewiesen. Hier waren die Mehrkosten jedoch vordergründig auf die kleinen Fallzahlen von «Kinder»-DRGs (also DRGs in denen ausschliesslich Kinder behandelt werden) zurückzuführen [29]. Zu berücksichtigen ist jedoch, dass sich alle empirischen Studien auf ältere Tarifversionen beziehen (bis einschliesslich Tarifversion 7.0) und seither zahlreiche Anpassungen stattgefunden haben, die zum Beispiel auf eine verbesserte Abbildung der Kosten von Outliern und Kindern abzielen (vgl. [36], [37]). Somit kann für neuere Tarifversionen tendenziell von einem Rückgang der Verzerrungen ausgegangen werden, was sich zum Beispiel in einer stärkeren Konzentration der Deckungsgrade pro DRG um 100% äussert [20].

Grundsätzlich wird diese Verzerrungsursache von uns als kritisch bewertet, weil auch künftig – trotz Evolution der Tarifstruktur – die dazu führenden Probleme nicht vollständig gelöst werden können. So ist aufgrund der Populationsgrösse in der Schweiz vor allem bei kleineren Fallgruppen weiterhin mit Verzerrungen zu rechnen. Auch das Outlier-Problem wird wohl nie vollständig gelöst werden können, wie langjährige, internationale Erfahrungen zeigen [26]. Unschärfen bei der Zuordnung der Fälle und Mängel bei der Leistungs- und Kostenerfassung sollten jedoch bei Verwendung einheitlicher Definitionen und Abgrenzungsmethoden zu weniger Problemen führen und mit zunehmender Systemerfahrung der Spitäler abnehmen.

Kosteninhomogenität innerhalb von DRGs

Kosteninhomogenität bemisst sich an der Streuung der Fallkosten innerhalb einer DRG. Fällt die Streuung hoch aus, so wird von einer Inhomogenität gesprochen. Kosteninhomogenität ist per se nicht problematisch. Sie wird allerdings zum Problem, wenn sich Fälle mit über- bzw. unterdurchschnittlichen Kosten bei einzelnen Spitälern häufen und diese Häufungen nicht auf Effizienzunterschiede der Spitäler zurückzuführen sind. In Folge können selbst für Spitäler mit gleichem Leistungsangebot aufgrund unterschiedlicher patientenbezogener Merkmale systematische Gewinne oder Verluste entstehen. Unterschiede im Leistungsangebot der Spitäler können die Entstehung von systematischen Gewinnen und Verlusten zusätzlich begünstigen, etwa, wenn Verbund- und Vorhalteleistungen , die einem Fall nicht spezifisch zuordenbar sind, vorliegen. Verbundleistungen sind Leistungen, die sich aus der inhaltlichen Vernetzung von Leistungen ergeben (d.h. es ist günstiger eine Kombination von Leistungen in einem Spital statt einzelne Leistungen in verschiedenen Spitälern zu erbringen). Vorhalteleistungen sind Leistungen, die die Sicherheit der Versorgung gewährleisten (z.B. Notfalldienste).

Patienten- und leistungsbezogene Unterschiede zwischen Spitälern, die zu unterschiedlichen Kostenstrukturen führen, wurden für die Schweiz mehrfach empirisch nachgewiesen (z.B. [5], [28], [29]). Zu den wichtigsten *patientenbezogenen Merkmalen* zählen gemäss diesen Studien vor allem Merkmale, welche die Komplexität und Art der Behandlung quantifizieren, wie zum Beispiel die Anzahl der Diagnosen eines Patienten. Aber auch der Gesundheitszustand vor Spitaleintritt, welcher zum Beispiel durch eine Aufnahme als Notfall oder die Überweisung aus einem anderen Spital gemessen wird, spielt eine wichtige Rolle. Zu den wichtigsten *leistungsbezogenen Merkmalen* zählen gemäss diesen Studien die Anteile der Hochdefizit- und Hochprofitfälle. Anders als erwartet zeigt sich, dass Leistungsdichte (gemessen durch die Anzahl angebotener DRGs) und Leistungsvolumen (z.B. gemessen durch die Anzahl der abgerechneten Patientenfälle) keinen Einfluss auf die Kosten haben, sofern für die Anteile der Hochdefizit- und Hochprofitfälle eines Spitals kontrolliert wird. Diese Studien zeigen, dass patienten- und leistungsbezogene Unterschiede zu einer systematischen Benachteiligung von Unispitälern [28], Kinderspitälern [29] und Endversorgerspitälern [26], aber auch Spezialkliniken [26], führen. Bei Letzteren sind vor allem pädiatrische Kliniken betroffen.

Im Zusammenhang mit der Kosteninhomogenität innerhalb von DRGs wurde insbesondere auch auf die Problematik der Hochkostenfälle verwiesen [5], [26]. Hochkostenfälle zeichnen sich durch die stark überdurchschnittlichen Kosten innerhalb einer Fallgruppe aus. Werden sie trotz Zusatzentgelten und Langliegerzuschlägen nicht entsprechend vergütet, werden sie zu Hochdefizitfällen. Diese können je nach Fallstruktur eines Spitals zu einer systematischen Benachteiligung führen. Dies zeigt sich insbesondere für Endversorgerspitäler (d.h. Spitäler, die am Ende der Versorgungskette stehen, wie bspw. Universitätsspitäler), aber auch für pädiatrische Spezialkliniken [26]. Systematische Vorteile ergeben sich hingegen für Regionalspitäler, welche sich durch einen deutlich tieferen Hochdefizitanteil bei gleichzeitig höherem Hochprofitanteil auszeichnen [26].

Die SwissDRG AG ist sich der Problematik von Hochkostenfällen durchaus bewusst [22] und hat mit Änderungen der Tarifstruktur auf die entsprechenden Studien reagiert (z.B. [20], [24], [25]). Diese Änderungen zielten vor allem darauf ab, hochkomplexe Fälle besser in der Tarifstruktur abzubilden, etwa durch eine aufwandsgerechtere Vergütung von Kombinationseingriffen, und spezielle Leistungsbereiche, wie zum Beispiel die Pädiatrie und Neonatologie oder die Intensivmedizin, aufwandsgerechter zu vergüten [36], [37]. Qualitative Auswertungen deuten darauf hin, dass Fortschritte in Bezug auf die Tarifstruktur bestehen und erfolgte Änderungen zumindest teilweise zu einer Entschärfung der Kosteninhomogenitätsproblematik führten [19]. Dennoch muss davon ausgegangen werden, dass dieses Problem langfristig nicht durch eine Evolution der Tarifstruktur

gelöst werden kann (vgl. z.B. auch [5], [30]). Denn eine gewisse Kosteninhomogenität ist jedem Fallpauschalensystem inhärent, und letztendlich auch gewünscht, um entsprechende Effizienzanreize zu setzen. Die Kosteninhomogenität innerhalb von DRGs ist deshalb als eine weitere relevante, wenn nicht sogar die wichtigste, Verzerrungsursache zu beurteilen.

2.4 Gewichtung im Spital Benchmarking

2.4.1 Motive und Voraussetzungen für eine Gewichtung

Von beiden in der Literatur identifizierten Gründen für eine Verzerrung in den Fallnormkosten (unvollständige Datengrundlage und Unsachgerechtigkeit der Fallnormkosten) gehen gegenwärtig Gewichtungsmotive aus. Grundsätzlich lassen sich diese Motive drei Ursachen zuordnen: Unit Nonresponse, Item Nonresponse und Messfehlern (vgl. Abbildung 4). Die ersten beiden Motive (Unit und Item Nonresponse) sind ausschliesslich auf Verzerrungen zurückzuführen, die von einer unvollständigen Datengrundlage ausgehen. Das letzte Motiv (Messfehler) ergibt sich aus Verzerrungen, die zu unsachgerechten Fallnormkosten führen.

Gemäss unserer Diskussion in 2.1.1 kämen in allen Fällen modellbasierte Gewichte zur Anwendung. Im Fall von Item Nonresponse sollte jedoch aus methodischen Gründen von einer Gewichtung abgesehen werden (siehe letzter Absatz auf Seite 8). Eine designbasierte Gewichtung kommt in unserem Fall nicht in Frage, da sich die Stichprobe der Netzwerkspitäler nicht aus einem konkreten Stichprobendesign ergibt, sondern aus der freiwilligen Teilnahme bzw. Datenlieferung der für SwissDRG relevanten akutsomatischen Spitäler.

Abbildung 4: Gewichtungsmotive im Spital Benchmarking

Gewichtungsmotive / Verzerrungsursachen		Designbasierte Gewichte	Modellbasierte Gewichte		
		Stichprobendesign basierte Verzerrungen	Unit Nonresponse	Item Nonresponse	Messfehler
Unvollständige Datengrundlage	Unzureichende Repräsentativität Netzwerkspitäler		- Keine Teilnahme - Keine Datenlieferung		
	Syst. Unterschiede nachträglich ausgeschlossener Spitäler/Fälle		- Verspätete Datenlieferung - Ausschluss von Spitälern	- Ausschluss v. Fällen wg. fehlenden Vollkosten, Abgrenzungsproblemen, möglichen Datenfehlern	
Unsachgerechte Fallnormkosten	Unzureichende Vergleichbarkeit der BM-relevanten Betriebskosten				- Abgrenzungs-probleme - Verwendung von Normabzügen
	Fehlbewertung von DRG Kostengewichten				- Tiefe DRG Fallzahlen - Kompressionseffekt - Codier Fehler
	Kosteninhomogenität innerhalb von DRGs				- System. Unterschiede in der Fallverteilung innerhalb einer DRG (z.B. durch leistungs- und patientenbez. Untersch.)

In der Praxis steht man modellbasierten Gewichten kritisch gegenüber (vgl. Ausführungen unter 2.1.2). Die Gründe sind methodische Schwierigkeiten, die im Zusammenhang mit der korrekten Ermittlung der Gewichte bestehen, und die mögliche Verschlimmerung der Verzerrung bei Anwendung «falscher» Gewichte. Deshalb

ist im Rahmen des Spital Benchmarkings auf jeden Fall von Gewichten abzuraten, wenn sich die Ursache für die Verzerrung anderweitig beheben lässt. Gemäss unserer Diskussion in 2.3 sind somit zwei Verzerrungsursachen im Rahmen der Gewichtungsfrage weiterhin relevant: die Fehlbewertung der DRG Kostengewichte und die Kosteninhomogenität innerhalb von DRGs. Beide Ursachen können langfristig nicht durch eine Evolution der Tarifstruktur gelöst werden. Das Gewichtungsmotiv wäre in beiden Fällen als Korrektur eines Messfehlers zu interpretieren, eine Anwendung, welche in der Praxis eher selten zu finden ist. Dies ist nicht zuletzt auch darauf zurückzuführen, dass bei diesem Gewichtungsmotiv etwaige Korrekturmöglichkeiten stark von der konkreten Anwendung abhängen (vgl. letzter Absatz Seite 9).

Damit eine Gewichtung im Rahmen des Spital Benchmarks zielführend wäre (d.h. zu einer korrekten Berechnung des Benchmark-Wertes führt), müssten alle Faktoren bekannt sein, die eine Verzerrung der Fallnormkosten herbeiführen. Einerseits wäre dies gegeben, wenn zumindest für einen Teil der Spitäler die Höhe des Messfehlers bekannt wäre und dieser korrekt auf die anderen Spitäler übertragen werden könnte (siehe erstes Anwendungsbeispiel im Fall von Messfehlern in 2.1.1). Dies ist hier jedoch nicht der Fall: Für kein Spital ist die Höhe des Messfehlers bekannt. Andererseits könnte die Höhe des Messfehlers für jedes Spital mittels des «Fallpauschalenmodells» von Polynomics geschätzt werden [5]. Dies würde jedoch voraussetzen, dass das «Fallpauschalenmodell», oder eine aktualisierte Version davon, vollständig und korrekt spezifiziert ist, d.h. dass alle Spitaleigenschaften, welche die Messfehler beeinflussen, im Modell berücksichtigt werden. Nur dann wären wir in der Lage, die Verzerrungen für alle Spitäler korrekt zu schätzen, womit theoretisch auch die Herleitung von adäquaten Gewichten möglich wäre.

Konkret ergibt sich bei einer Gewichtung im Rahmen des Spital Benchmarkings jedoch noch ein weiteres Problem. Es geht beim Spital Benchmarking nicht nur um die korrekte Berechnung des Benchmark-Wertes (was der klassischen Anwendung von Gewichten entspricht), sondern auch um eine Unterteilung in effiziente und ineffiziente Spitäler. Diese Unterteilung kann jedoch auf Basis der (unkorrigierten) Fallnormkosten nicht vorgenommen werden, weil sich durch eine Gewichtung der Spitäler die Fallnormkosten der Spitäler und somit die Reihenfolge der Spitäler nicht verändern. In Folge können Spitäler, die eigentlich ineffizient sind, als effizient beurteilt werden, und umgekehrt, d.h. das ursprüngliche Problem bleibt bestehen. Eine korrekte Unterteilung in effiziente und ineffiziente Spitäler kann ausschliesslich über eine Korrektur der Fallnormkosten erreicht werden, mit anschliessendem Benchmarking auf Basis der korrigierten Fallnormkosten. Abbildung 5 verdeutlicht diese Problematik anhand eines Beispiels mit 10 fiktiven Spitälern.

2.4.2 Gewichtung im Spital Benchmarking anhand eines Beispiels

Grafik A in Abbildung 5 zeigt die Fallnormkosten, der Höhe nach absteigend geordnet, für zehn fiktive Spitäler. Die höchsten Fallnormkosten weist das Spital 1 auf (14'500 CHF), die tiefsten das Spital 10 (8’480 CHF). Wird der Benchmark beim 25. Perzentil angesetzt, beträgt der «effiziente Preis» (Baserate) 9'288 CHF. Die Spitäler 8, 9 und 10 würden unter der naiven Annahme, dass die Fallnormkosten der Spitäler vergleichbar sind, d.h. die Unterschiede in den Fallnormkosten der Spitäler ausschliesslich auf Effizienzunterschiede zurückzuführen sind, als effizient klassifiziert, die restlichen nicht. Dieses Beispiel entspricht dem derzeit durchgeführten Spital Benchmarking ohne Gewichtung oder Korrektur der Fallnormkosten. Wie in 2.3 dargelegt, sind die Fallnormkosten der Spitäler aber nicht direkt vergleichbar, weil sie nicht nur Unterschiede in der Effizienz der Spitäler, sondern auch Unterschiede in der Patienten- und Leistungsstruktur der Spitäler abbilden.

Grafik B zeigt die Fallnormkosten plus eine entsprechende «Korrektur». Die «Korrektur» enthält alle gerechtfertigten Kostenunterschiede, die nicht über die SwissDRG Kostengewichte abgebildet werden, und folglich

auch nicht über die Normierung der Fallkosten mit dem Case Mix Index herausgerechnet werden. Gerechtfertigte Unterschiede können zum Beispiel durch besondere Verbund- und Vorhalteleistungen im Rahmen von Leistungsaufträgen, regional bedingte Kostenunterschiede (z.B. Löhne) oder unterschiedliche spitalspezifische Patientenmerkmale (z.B. Kinder vs. Erwachsene) entstehen. Wir modellieren die «Korrektur» mittels normalverteilter Zufallsvariable mit Mittelwert 10 und Standardabweichung 400. Dabei orientieren wir uns an den Resultaten der Studie von Polynomics, welche auf keinen systematischen Zusammenhang zwischen der Höhe der Fallnormkosten gemäss SwissDRG und den «korrigierten Fallnormkosten» schliessen lässt und erhalten «Korrekturen» in ähnlicher Grössenordnung (vgl. Abb.2 in [5]). Bei den Spitälern 1, 4 und 6 ist die «Korrektur» negativ, d.h. die «korrigierten Fallnormkosten» sind tiefer als die beobachteten. Bei diesen Spitälern könnte es sich beispielsweise um Spezialkliniken mit eingeschränktem Leistungsangebot handeln, die besonders kranke Patienten behandeln (viele High-Outlier). Es könnte sich aber auch um Universitätsspitaler handeln, deren breites Leistungsangebot in Kombination mit vielen komplexen Fällen, hohe Verbund- und Vorhalteleistungen erfordert. Bei den übrigen Spitälern in unserem Beispiel liegen die «korrigierten» Fallnormkosten über den beobachteten. Dabei könnte es sich zum Beispiel um Regional- oder Spezialkliniken handeln, die weniger komplexe Patienten behandeln und/oder deren eingeschränktes Leistungsangebot weniger Verbund- und Vorhalteleistungen erfordert.

Grafik C in Abbildung 5 zeigt die geordneten «korrigierten Fallnormkosten». Ein Vergleich der «korrigierten Fallnormkosten» lässt nun Rückschlüsse auf die Effizienz der Spitäler zu. Der «korrekte» Benchmark liegt bei 9'400 CHF. Die Spitäler 4, 6 und 10 erweisen sich als die tatsächlich effizienten Spitäler. Beim naiven Vergleich der (unkorrigierten) Fallnormkosten in *Grafik A* wurde jedoch nur Spital 10 als effizient identifiziert. Die Spitäler 4 und 6 wurden zu Unrecht als ineffizient beurteilt. Die Spitäler 8 und 9 hingegen, welche gemessen an den (unkorrigierten) Fallnormkosten in *Grafik A* fälschlicherweise als effizient beurteilt wurden, liegen in Wahrheit über dem Benchmark, d.h. sie sind ineffizient.

Grafik D zeigt zum Vergleich das Resultat des Benchmarkings auf Basis der gewichteten Fallnormkosten. Die Gewichte wurden so generiert, dass sie die Verzerrung im Benchmark vollständig korrigieren, d.h. also den Benchmark-Wert korrekt berechnen. So beträgt das gewichtete 25. Perzentil der unkorrigierten Fallnormkosten ebenfalls 9’400 CHF. Weil sich jedoch die Reihenfolge der Spitäler durch die Gewichtung nicht ändert, bzw. beim Vergleich der Fallnormkosten mit dem Benchmark auf die unkorrigierten Fallnormkosten zurückgegriffen wird, lassen sich keine Effizienzaussagen treffen.

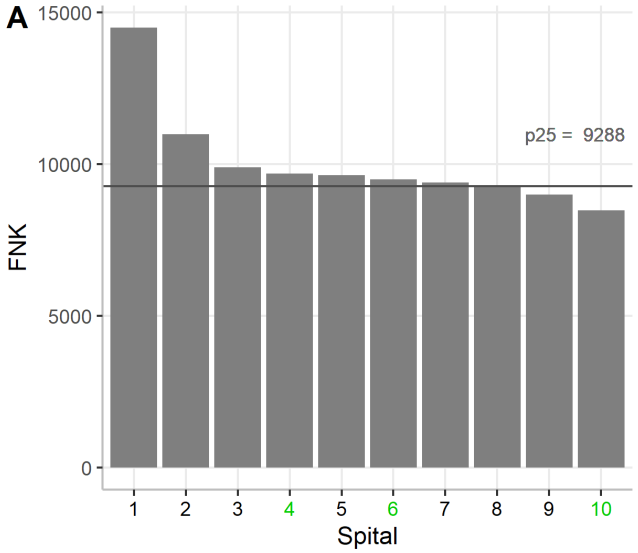
Ein Benchmarking mittels gewichteten Fallnormkosten ist demnach nicht zielführend, selbst wenn die Gewichte korrekt sind (d.h., wenn sie vollständig für die gerechtfertigten Kostenunterschiede korrigieren). Ferner muss davon ausgegangen werden, dass eine Gewichtung nicht einmal zur korrekten Ermittlung des Benchmark-Wertes führt, da in der Regel inkorrekte Gewichte verwendet werden. Dies wird aller Voraussicht nach vor allem dann der Fall sein, wenn auf Adhoc-Gewichte, wie beispielsweise den Case Mix oder die Anzahl Fälle pro Spital (Grösse des Spitals) zurückgegriffen wird, welche gemäss empirischen Studien die Unterschiede in den Fallnormkosten gar nicht oder nur teilweise zu erklären vermögen (z.B. [5]). Insbesondere eine Gewichtung gemäss Anzahl Fälle der Spitäler läuft der Idee des Benchmarkings zuwider. Der Grund ist, dass es beim Benchmarking um einen Effizienzvergleich aller Spitäler auf Ebene Betrieb geht und mögliche Ineffizienzen, die zum Beispiel aufgrund fehlender Skaleneffekte entstehen können, herunterskaliert bzw. verwässert werden.

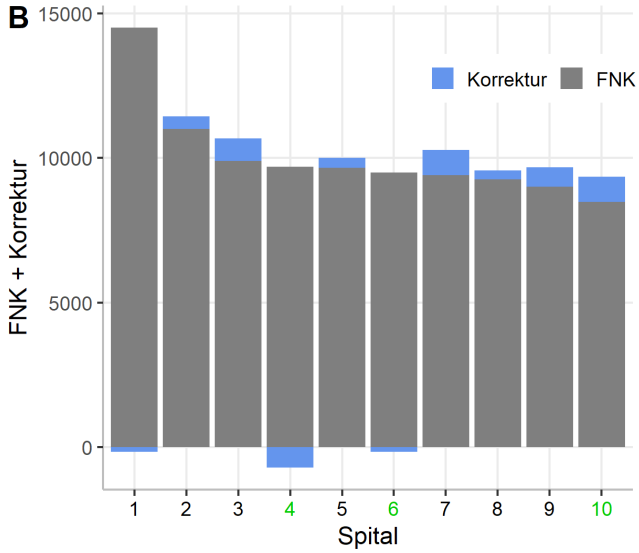
Abbildung 5: Gewichtung im Spital Benchmarking am Beispiel von 10 fiktiven Spitälern

A zeigt die Fallnormkosten für zehn fiktive Spitäler.

Das 25. Perzentil (Benchmark) liegt bei 9'288 CHF.

Grün markiert sind jene Spitäler, die eigentlich (gemessen an den korrigierten Fallnormkosten) effizient sind, d.h. unter dem Benchmark liegen (vgl. **C**).





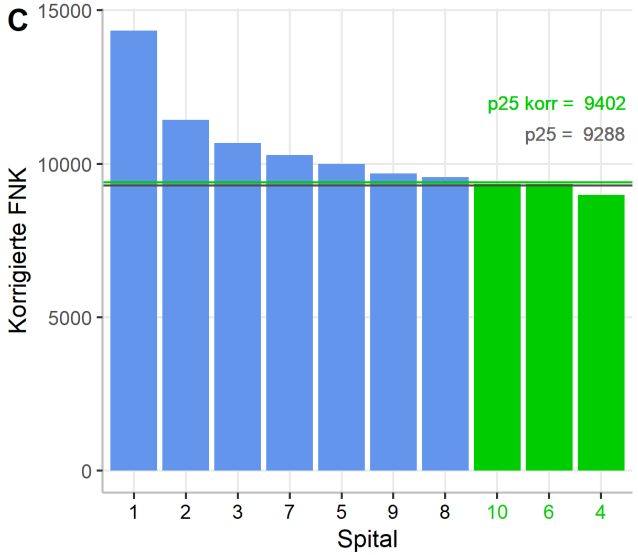
B zeigt die Fallnormkosten plus Korrektur.

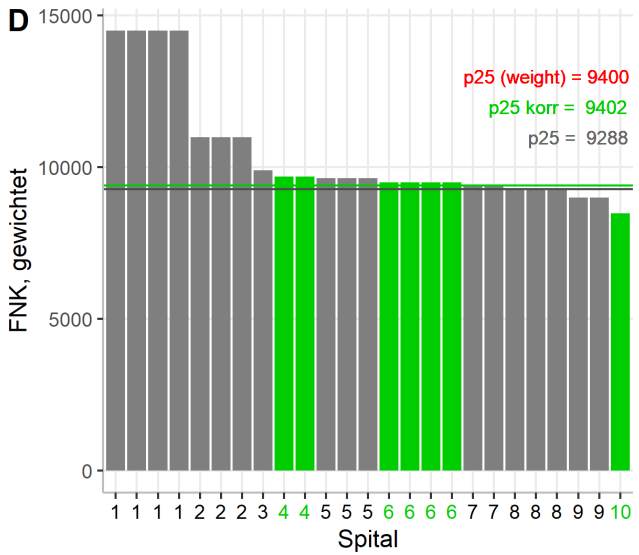
Ausgehend davon, dass die Höhe der Verzerrung nicht systematisch mit der Höhe der FNK zusammenzuhängen scheint [5], wurde die Korrektur mittels normalverteilter Zufallsvariable generiert (μ=10, σ=400).

C zeigt die korrigierten Fallnormkosten.

Es zeigt sich, dass die Spitäler 4 und 6 eigentlich effizient sind, wohingegen die Spitäler 8 und 9 eigentlich ineffizient sind.

Das 25. Perzentil der korrekten Fallnormkosten liegt bei gut 9'400 CHF.





D zeigt eine Möglichkeit der «Korrektur» des 25. Perzentils der Fallnormkosten mittels Gewichtung.

Das gewichtete 25. Perz. der FNK liegt nun bei 9'400 CHF.

Ein Spital Benchmarking ist aufgrund der unveränderten Reihenfolge aber nicht möglich bzw. fällt nicht korrekt aus.

3 Fazit

Die vorliegende Studie erarbeitet die Motive für und beurteilt die Zweckmässigkeit von einer Gewichtung im Rahmen des Spital Benchmarkings. Sie kommt zu vier zentralen Ergebnissen:

Zahlreiche Gewichtungsmotive

Es gibt derzeit verschiedene Ursachen für eine Verzerrung der Fallnormkosten, die eine Gewichtung im Rahmen des Spital Benchmarkings denkbar machen. Diese Ursachen unterscheiden sich jedoch hinsichtlich ihrer Relevanz. Verzerrungen, die auf eine unvollständige Datengrundlage oder auf die inkorrekte und uneinheitliche Berechnung der benchmarkrelevanten Betriebskosten zurückzuführen sind, lassen sich langfristig beheben (z.B. durch vollständigere Daten und methodische Verbesserungen) und werden deshalb als weniger relevant eingestuft. Verzerrungen, die hingegen auf Fehlbewertungen von DRG Kostengewichten (z.B. wegen kleiner Fallgruppen oder Outlier-Korrekturen) und Kosteninhomogenitäten innerhalb von DRGs (z.B. wegen leistungs- und patientenbezogenen Unterschieden zwischen den Spitälern) zurückzuführen sind, lassen sich nicht vollständig durch die Evolution der heutigen Tarifstruktur beheben. Sie sind system-inhärent und daher auch langfristig von Bedeutung.

Ausschliesslich modellbasierte Gewichte

Ausgehend von den identifizierten Verzerrungsursachen kämen ausschliesslich modellbasierte Gewichte zur Anwendung. Diese Gewichte werden im Allgemeinen kritisch beurteilt, da sie basierend auf sehr starken Annahmen, die nicht überprüft werden können, geschätzt werden müssen, und es bei Verwendung «inkorrekter» Gewichte zu einer Verschlimmerung der Verzerrung kommen kann. **Deshalb ist im Rahmen des Spital Benchmarkings auf jeden Fall von Gewichten abzuraten, wenn sich die Ursache für die Verzerrung anderweitig beheben lässt**. Konkret heisst das, dass nur von zwei Verzerrungsursachen (Fehlbewertungen von DRG Kostengewichten, Kosteninhomogenitäten innerhalb von DRGs) ein begründetes Gewichtungsmotiv ausgeht.

Unzweckmässigkeit der Gewichtung im Spital Benchmarking

Die Anwendung von Gewichten im Rahmen des Spital Benchmarkings ist in keinem Fall zweckmässig. Zwar könnte bei Verwendung von korrekten Gewichten der Benchmark-Wert korrekt berechnet werden. Eine Unterteilung in effiziente und ineffiziente Spitäler ist aber dennoch nicht möglich, weil sich durch die Gewichtung der Spitäler die Fallnormkosten und somit die Reihenfolge der Spitäler nicht verändern.

Korrektur der Fallnormkosten erforderlich

Eine korrekte Unterteilung in effiziente und ineffiziente Spitäler kann ausschliesslich über eine Korrektur der Fallnormkosten erreicht werden. Die Korrektur der Fallnormkosten erfolgt dabei idealerweise anhand eines empirischen Modells, wie z.B. dem «Fallpauschalenmodell» von Polynomics [5]. Die auf diese Weise um die gerechtfertigten Unterschiede korrigierten Fallnormkosten erlauben ein valides schweizweites Benchmarking. Es müssten dann zwar immer noch differenzierte Baserates verhandelt werden, weil die Korrektur nachgelagert an die Berechnung der Kostengewichte durch SwissDRG erfolgt. Die Höhe der Verzerrung der Fallnormkosten würde aber eine valide Grundlage für die Verhandlungen bilden.

Alternativ wäre eine Art «gerechtfertigter» Lastenausgleich zwischen den Spitälern denkbar, der nachträglich für die Verzerrungen in den Fallnormkosten kompensiert. Dies hätte den Vorteil, dass es nur noch eine schweizweit einheitlich geltende Baserate gäbe und somit keine Verhandlungen mehr nötig wären. Auch diese Lösung setzt die Korrektur (Sachgerechtigkeit) der Fallnormkosten voraus, um die Höhe etwaiger Ausgleichszahlungen zwischen den Spitälern ex-post zu bestimmen.

Literaturverzeichnis

[1] BVGer-Urteil und Bundesverfassungsgericht, C-1698/2013; Auszug aus dem Urteil der Abteilung III i.S. 46 Krankenversicherer gegen Luzerner Kantonsspital und Regierungsrat des Kantons Luzern C–1698/2013. 2014.

[2] S. Iseli, M. Fierri Kovacs, M. Trüb, und M. Jung, Spitaltarife - Praxis des Preisüberwachers bei der Prüfung von akut-stationären Spitaltarifen. 2016.

[3] S. Spika und H. Keune, „Lösungsansätze für eine faire Vergütung“, Competence, Bd. 3, S. 22–23, 2016.

[4] M. Waldner, „Neue Spitalfinanzierung am Scheideweg. Eine neue Studie lässt Korrekturen bei der Benchmarking-Methodik dringend angezeigt erscheinen“, Jusletter, Nr. 23, 2015.

[5] P. Widmer, M. Trottmann, und H. Telser, „Das Fallpauschalenmodell: Leistungsbezogene Basispreise unter SwissDRG“, 2015.

[6] BAG, „Publikation von «schweregradbereinigten Fallkosten» im Rahmen von Artikel 49 Absatz 8 KVG - Konzept“, 2019.

[7] S. Gabler, J. H. Hoffmeyer-Zlotnik, und D. Krebs, Gewichtung in der Umfragepraxis. Springer, 1994.

[8] M. Bertolet, „To Weight or Not to Weight: Incorporating Sampling Effects into Model-Based Survey Analysis“, Ph. D. Dissertation. Department of Statistics, Carnegie Mellon University, 2008.

[9] BFS Bundesamt für Statistik, „Die Schweizerische Gesundheitsbefragung 2017 in Kürze. Konzept, Methode, Durchführung.“ 2018.

[10] N. Baur und J. Blasius, Handbuch Methoden der empirischen Sozialforschung. Springer, 2014.

[11] G. B. Durrand und C. Skinner, „Using Missing Data Methods to Correct for Measurement Error in a Distribution Function“, Surv. Methodol., Bd. 32, Nr. 1, S. 25–36, 2006.

[12] J. M. Brick, „Unit nonresponse and weighting adjustments: A critical review“. Versita, 2013.

[13] R. J. Little, „Survey nonresponse adjustments for estimates of means“, Int. Stat. Rev. Int. Stat., S. 139–157, 1986.

[14] J. M. Wooldridge, „Inverse probability weighted estimation for general missing data problems“, J. Econom., Bd. 141, Nr. 2, S. 1281–1301, 2007.

[15] D. Loveridge, L. Georghiou, und M. Nedeva, United Kingdom Techology Foresight Programme: Delphi Survey. HM Stationery Office, 1995.

[16] G. Solon, S. J. Haider, und J. M. Wooldridge, „What are we weighting for?“, J. Hum. Resour., Bd. 50, Nr. 2, S. 301–316, 2015.

[17] P. R. Rosenbaum und D. B. Rubin, „The central role of the propensity score in observational studies for causal effects“, Biometrika, Bd. 70, Nr. 1, S. 41–55, 1983.

[18] A. Gelman, „Struggles with survey weighting and regression modeling“, Stat. Sci., Bd. 22, Nr. 2, S. 153–164, 2007.

[19] M. Lobsiger und M. Frey, „Evaluation der KVG-Revision im Bereich der Spitalfinanzierung. Auswirkungen der Revision auf die Kosten und die Finanzierung des Versorgungssystems Schlussbericht.“, 2019.

[20] SwissDRG, „Änderungen in SwissDRG Version 9.0 gegenüber Version 8.0“, 2019.

[21] SwissDRG, „Einführung für die neuen Netzwerkspitäler SwissDRG. Erhebung der Falldaten der Schweizer Spitäler im akutsomatischen Bereich.“ 2019.

[22] SwissDRG, „8. Informationsveranstaltung SwissDRG - Version 5.0/2016“, 2015.

[23] Bundesrat, Verordnung über die Krankenversicherung (KVV). Änderungsvorschlag des Bundesrates. 2020.

[24] SwissDRG, „Bericht zur Weiterentwicklung der SwissDRG Tarifstruktur 7.0/2018“, SwissDRG, 2018.

[25] SwissDRG, „Änderungen in SwissDRG Version 8.0 gegenüber Version 7.0“, SwissDRG, 2018.

[26] P. Hochuli, P. Widmer, und H. Telser, „Faire Abgeltung von Hochkostenfällen in DRG-Systemen – Internationale Erfahrungen und Lösungskonzepte“, 2017.

[27] S. Spika und H. Keune, „Benchmarking über die Fallnormkosten–bitte Gleiches mit Gleichem!“, Bull. Médecins Suisses, Bd. 95, Nr. 12, 2014.

[28] P. Widmer, S. Spika, und H. Telser, „Leistungsorientierte Vergütung mit dem Fallpauschalensystem SwissDRG. Gleicher Preis für gleiche Leistung?“, 2015.

[29] P. Widmer, P. Hochuli, und H. Telser, „Theoretische und empirische Analyse der Mehrkosten der Kinderspitäler unter SwissDRG“, 2017.
[30] M. Waldner, „Tarifstruktur SwissDRG im Jahr 2018: Basis für einen fairen Wettbewerb?“, Jubil. zum SwissDRG Forum 2018, 2018.
[31] M. Waldner, „Anlagenutzungskosten und neue Spitalfinanzierung. Die geltende VKL gefährdet eine KVG-konforme Anlagenbewertung.“, Aktuelle Juristische Prax. AJP, Nr. 10, S. 1244–1251, 2017.
[32] BVGer-Urteil, C-2283/2013 und C-3617/2013 Auszug aus dem (Teil-)Urteil der Abteilung III i.S. Stadt Zürich. 2014.
[33] P. Widmer, „SwissDRG: Ein Vergütungssystem mit ungleichen finanziellen Risiken für die Spitäler?“, Unternehm., Bd. 70, Nr. 3, S. 210–226, 2016.
[34] SwissDRG, „Editorial zur Studie: Theoretische und empirische Analyse zu den Mehrkosten der Kinderspitäler unter SwissDRG. Studie im Auftrag der SwissDRG AG in Zusammenarbeit mit AllKidS“, 2017.
[35] P. Widmer, H. Telser, und T. Uebelhart, „Die Spitalversorgung im Spannungsfeld der kantonalen Spitalpolitik. Aktualisierung 2015“, 2016.
[36] SwissDRG, „Entwicklungsschwerpunkte und Abbildung von speziellen Leistungsbereichen in der SwissDRG-Version 8.0“, 2019.
[37] SwissDRG, „Abbildung spezieller Leistungsbereiche und Entwicklungsschwerpunkte der SwissDRG Version 9.0“, 2019.