POLYNOMICS

Weiterentwicklung der statistischen Methode zur Prüfung der Wirtschaftlichkeit

Schlussbericht

Studie im Auftrag von FMH, santésuisse und curafutura

Weiterentwicklung der statistischen Methoden zur Überprüfung der Wirtschaftlichkeit

Studie im Auftrag von FMH, santésuisse und curafutura

Dr. Maria Trottmann, Barbara Fischer, Dr. Tobias von Rechenberg, Dr. Harry Telser

September 2017

Verdankungen
Wir danken Prof. Dr. Stefan Boes (Direktor des Zentrums für Gesundheit, Politik und Ökonomie an der Universität Luzern, Leiter des SNF-Programms Swiss Learning Health Systems) für die sorgfältige Lektüre und exzellenten Kommentare zu einer früheren Version dieses Berichts.
Ebenfalls danken wir der Begleitgruppe der Auftraggeber bestehend aus Marc Bill (santésuisse), Mirjam D'Angelo (santésuisse), Dr. med. Andreas Häfeli (FMH), Thomas Kessler (FMH), Dr. Philip Moline (NewIndex) und Anke Trittin (curafutura), deren fundierte Rückmeldungen zu wesentlichen Verbesserungen der Studie geführt haben.

Inhaltsverzeichnis

Abkürzungsverzeichnis

Adj. R^2	Adjusted R^2; korrigiertes Bestimmtheitsmass
AGG	Alters- und Geschlechtsgruppe
AIC	Akaike's Information Criterion; Akaikes Informationskriterium
ANOVA	Analysis of Variance; Varianzanalyse
BIC	Bayesian Information Criterion; bayesianisches Informationskriterium
COPD	Chronic Obstructive Pulmonary Disease; chronisch obstruktive Lungenerkrankung
DDD	Defined Daily Dosis; definierte täglich Dosis eines Wirkstoffes
GLM	Generalized Linear Model; verallgemeinerte lineare Modelle
MAPE	Mean Absolute Prediction Error; durchschnittlicher absoluter Prognosefehler
MiGeL	Mittel und Gegenständeliste
N	Anzahl Beobachtungen
OLS	Ordinary Least Squares; Methode der kleinsten Quadrate
PCG	Pharmaceutical Cost Group; pharmazeutische Kostengruppe
PE	Praxisspezifischer Effekt
PPV	Positive Predictive Value; positiver Vorhersagewert
RSS	Residual Sum of Squares; Residuenquadratsumme
SCD	Standardized Cost Difference; standardisiertes Kostenverhältnis zwischen beobachteten und erwarteten Kosten einer Arztpraxis
Std. Abw.	Standardabweichung
Tarmed	Tarif zur Vergütung ambulanter ärztlicher Leistungen in der Schweiz
ZSR	Zahlstellenregister der Sasis AG

1 In Kürze

1.1 Ausgangslage

Gemäss Art. 56 Abs. 6 KVG müssen sich die Versicherer und die Leistungserbringer auf ein Verfahren zur Wirtschaftlichkeitsprüfung einigen. Im ersten Schritt dieses Verfahrens werden mit statistischen Methoden Praxen identifiziert, deren Kosten unter Berücksichtigung des Patientenstamms deutlich über den durchschnittlichen Kosten der Facharztgruppe liegen. Diese Praxen werden im weiteren Verfahren einzeln überprüft.

Seit 2004 kommt für dieses statistische Screening die sogenannte ANOVA-Methode zum Einsatz (Roth und Stahel 2005, Kaiser 2016). Grob gesprochen werden auf einer ersten Stufe die logarithmierten mittleren Kosten pro Arzt um den Effekt der Alters- und Geschlechtsgruppe (im Folgenden: AGG) seiner Patienten bereinigt. In einer zweiten Stufe folgt die Bereinigung um den Einfluss der Facharztgruppe und des Kantons. Die vorliegende Studie hat das Ziel, die bisherige Methode weiterzuentwickeln, insbesondere durch eine verstärkte Berücksichtigung der Morbidität des Patientenstamms. Dabei stehen insbesondere folgende Fragen im Vordergrund:

1. Diskussion des Schätzmodells im Lichte der internationalen Fachliteratur: Ist die durch Kaiser (2016) vorgeschlagene Spezifikation zweckmässig, beziehungsweise wie könnte die Spezifikation erweitert werden?
2. Welche zusätzlichen Morbiditätsfaktoren eigenen sich für den Einbezug in das Modell? Wie können diese Morbiditätsfaktoren empirisch umgesetzt werden?
3. Welche zusätzlichen Charakteristika der Arztpraxis, insbesondere der Praxisstandort, haben einen erheblichen Einfluss auf die Kosten pro Praxis? Wie können diese in das Modell eingefügt werden?
4. Welche Treffgenauigkeit (Anzahl falsch positiv bzw. falsch negativ klassifizierte Praxen) wird mit dem statistischen Screening erreicht? Welche Erweiterungen des Schätzmodells könnten zu einer verbesserten Treffergenauigkeit führen?
5. Könnte durch eine Berechnung mit patientenbezogenen Daten eine wesentliche Verbesserung der Treffergenauigkeit erreicht werden?

1.2 Beurteilung und Erweiterungen des Schätzmodells

1.2.1 Das zweistufige Schätzmodell

Kaiser (2016) schlägt vor, auf der ersten Stufe ein Fixed-Effects-Modell zu schätzen. Dies erlaubt die Berechnung eines praxisspezifischen Effekts, welcher um den Einfluss des Patientenstamms bereinigt ist. Intuitiv gesprochen gibt der Praxiseffekt an, inwiefern die Praxis durchschnittlich von den Kosten abweicht, welche für eine Praxis der gleichen Facharztgruppe und dem gleichen Patientenkollektiv erwartet würden. Aus unserer Sicht ist diese Spezifikation zweckmässig. Bei der Interpretation des Praxiseffektes muss jedoch beachtet werden, dass ein hoher Wert nicht zwingend als Ineffizienz interpretiert werden kann. So könnten beispielsweise auch Besonderheiten im Leistungsspektrum vorliegen, welche dazu führen, dass die Kosten einer Praxis vom Durchschnitt der Facharztgruppe abweichen. Solche systematischen Abweichungen kann das statistische Modell nicht korrigieren, ausser sie sind in den berücksichtigten Morbiditätsfaktoren reflektiert.

In einem traditionellen Fixed-Effects-Modell, wie es für die erste Stufe vorgeschlagen wurde, können keine Faktoren berücksichtigt werden, die pro Praxis konstant sind. Kaiser (2016) schlägt daher vor, um Faktoren wie beispielsweise den Praxisstandort auf der zweiten Stufe zu korrigieren. Wir halten diese zweistufige Schätzung für ein valides, robustes Verfahren und für die beste Lösung in der vorliegenden Frage. Alternative Spezifikationen des Fixed-Effects-Modells existieren in der Fachliteratur, sie sind aus unserer Sicht aber noch nicht ausreichend getestet und etabliert, als dass wir eine operative Umsetzung in der Wirtschaftlichkeitsprüfung empfehlen. Aus den bereinigten Praxiseffekten wird anschliessend ein Index berechnet, der für jede Praxis anzeigt, wie viele Prozent sie über oder unter den erwarteten Kosten liegt.

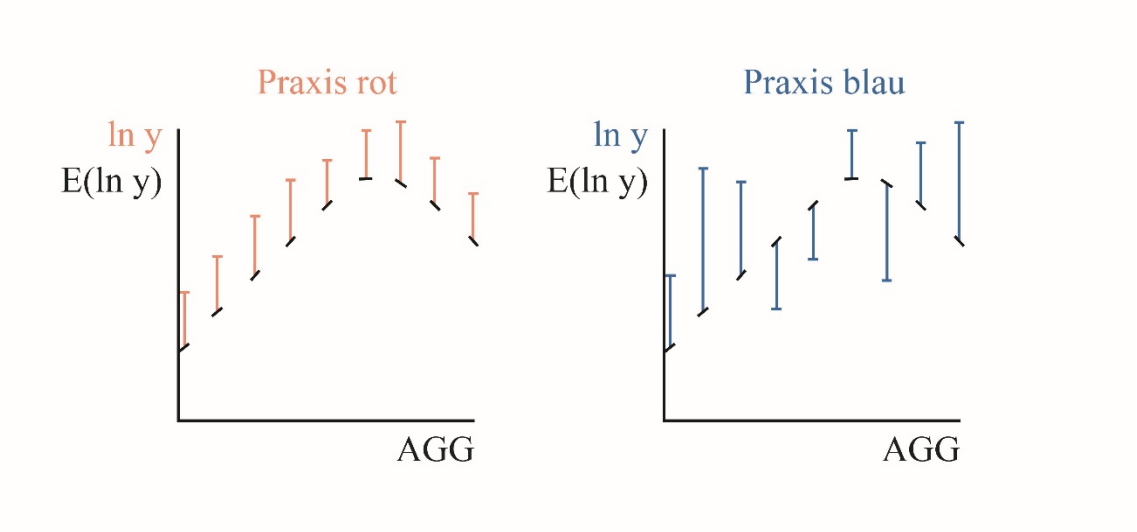
1.2.2 Empfohlene Erweiterung um einen Unsicherheitsfaktor

Wir empfehlen zusätzlich zu berücksichtigen, dass der praxisspezifische Effekt nicht punktgenau, sondern nur mit einer gewissen statistischen Unsicherheit ermittelt werden kann. Dazu berechnen wir einen Unsicherheitsindikator (ähnlich einer Standardabweichung). Intuitiv kann der Unsicherheitsindikator folgendermassen interpretiert werden: Weicht eine Praxis bei all ihren Beobachtungen im ähnlichen Umfang von den durch das Modell vorhergesagten Kosten ab, ist der Unsicherheitsfaktor gering. Weichen aber beispielsweise einige Beobachtungen sehr stark positiv ab, andere jedoch kaum oder negativ, ist der Unsicherheitsindikator hoch.

In Abbildung 1 wird dies anhand zweier fiktiver Beispielspraxen illustriert: Die schwarzen Punkte stellen die Regressionsergebnisse aus dem statistischen Modell dar (= erwartete Kostenwerte einer Praxis bei gegebenen Morbiditätsvariablen). Die senkrechten farbigen Linien sind die Abweichungen der beobachteten Werte vom erwarteten Wert. Der praxisspezifische Effekt ist jeweils der Durchschnitt der Abweichungen (= mittlere Länge der farbigen Linien). Praxis «rot» weicht bei allen Beobachtungen etwa gleich stark vom erwarteten Wert ab. Bei ihr ist die Streuung und somit der Unsicherheitsindikator gering. Anders ist die Situation bei Praxis «blau»: Ihre Abweichungen schwanken stark und haben nicht alle das gleiche Vorzeichen. Der Unsicherheitsindikator ist bei Praxis «blau» wesentlich höher als bei Praxis «rot». Es ist aber möglich, dass die durchschnittliche Abweichung und somit der praxisspezifische Effekt in beiden Praxen gleich sind.

Mit dem Unsicherheitsindikator soll das Problem adressiert werden, dass hohe Ausreisserwerte einen praxisspezifischen Effekt verändern können, obwohl sie höchstwahrscheinlich nichts mit der Wirtschaftlichkeit einer Praxis zu tun haben. Solche Ausreisserwerte werden pro Praxis nur wenige Beobachtungen betreffen. Falls bei einer Praxis nur wenige Werte – dafür stark – über den erwarteten Kosten liegen, sollte dies daher anders beurteilt werden als bei einer Praxis, wo alle Beobachtungen systematisch über den erwarteten Kosten liegen. Statistisch gesprochen kann der praxisspezifische Effekt mit grösserer Sicherheit identifiziert werden, falls die Varianz der Abweichungen vom erwarteten Wert (Residuen) innerhalb einer Arztpraxis vergleichsweise gering ist.

Abbildung 1 Illustration des Unsicherheitsfaktors



Quelle: Eigene Darstellung, Polynomics.

1.3 Einbezug weiterer Morbiditätsindikatoren

1.3.1 Auswahl und empirische Umsetzung der Morbiditätsindikatoren

In der internationalen Fachliteratur wird bei allen Verfahren zum sogenannten «Physician Profiling» die Morbidität des Patientenstamms berücksichtigt. Auch für die Schweiz wurde eine solche Berücksichtigung wiederholt gefordert (Schwenkglenks 2010; Wasem, Lux und Dahl 2010). Das Hauptziel des Projekts war daher eine bessere Morbiditätskorrektur. In einer Voranalyse haben wir mögliche Morbiditätsindikatoren nach den drei Kriterien «erwarteter Erklärungsgehalt», «Exogenität» und «Datenverfügbarkeit» beurteilt. Daraus wurden folgende Indikatoren zur datenbasierten Prüfung ausgewählt: Wahlfranchisen, Spital-im-Vorjahr und pharmazeutische Kostengruppen (PCG). In der Fachliteratur sind zusätzlich diagnosebezogene Indikatoren stark verbreitet. In der Schweiz erfüllen sie jedoch das Kriterium der «Datenverfügbarkeit» nicht, so dass ein empirischer Test nicht in Frage kam.

Das Ziel war, die Morbiditätsindikatoren so zu spezifizieren, dass sie sowohl mit patientenbezogenen als auch mit aggregierten Daten gebildet werden können. Die aggregierten Daten (aggregiert nach AGG) stammen dabei aus dem Daten- bzw. Tarifpool der Sasis AG. Sie sind aktuell die einzige Datenquelle in der Schweiz, in welcher die durch die obligatorische Krankenversicherung bezahlten ambulanten Leistungen flächendeckend erfasst werden. Die patientenbezogenen Daten wurden anonymisiert von drei Versicherern zur Verfügung gestellt.

- **Umsetzung des Indikators «Wahlfranchisen»**
  Die Patienten in der Schweiz können zwischen sechs Franchisestufen wählen. Für die Analyse haben wir sie in zwei Gruppen zusammengefasst: Als niedrig gilt die ordentliche Franchise und die erste Wahlfranchise (bei Erwachsenen CHF 300 und CHF 500), als hoch alle höheren Franchisestufen. Diese Gliederung in niedrig und hoch ist in der Literatur verbreitet. Eine detaillierte Gliederung mit weiteren Franchisestufen erhöht die Modellgüte kaum. Auf der Patientenebene ist die Franchisestufe ein [0/1]-Indikator. Auf der Praxisebene setzen wir den Indikator als Anteil der Patienten mit einer hohen Franchisestufe pro AGG und Praxis um.

- **Umsetzung des Indikators «Spital-im-Vorjahr»**
  Die Variable «Spital-im-Vorjahr» gibt an, ob der Patient im Vorjahr einen Spitalaufenthalt hatte. Sie ist ebenfalls ein [0/1]-Indikator auf Patientenebene. Auf der Praxisebene spezifizieren wir sie als Anteil pro AGG und Praxis.

- **Umsetzung des Indikators «Pharmazeutische Kostengruppen» («PCG»)**
  Die Idee hinter pharmazeutischen Kostengruppen ist, Medikamentenabrechnungen zu nutzen, um Morbiditätsindikatoren für die statistische Kostenprognose zu bilden. Das wichtigste Element zur Bildung von PCG ist ein Klassifikationsverfahren, welches Wirkstoffe Indikationsgebieten zuteilt. Wir nutzen hier ein Klassifikationsverfahren, welches im Auftrag des BAG für den Risikoausgleich zwischen den Krankenversicherern entwickelt wurde (Trottmann et al. 2015).
  Auch wenn ein eigenes Klassifikationsverfahren die Modellgüte möglicherweise erhöhen würde, ist es auch in Zukunft eine attraktive Option, die Klassifikation des Risikoausgleichs für die Wirtschaftlichkeitsverfahren zu verwenden. Diese wird durch das BAG kontinuierlich weitergepflegt und an neue Entwicklungen in der Arzneimittelversorgung angepasst.
  Die Liste des BAG enthält aktuell 24 PCG. Aus unserer Sicht sollten aber nicht alle PCG für alle Facharztgruppen berücksichtigt werden. Konkret empfehlen wir, eine PCG nur dann zu berücksichtigen, wenn innerhalb einer Facharztgruppe durch über 30 Praxen eine Mindestmenge an PCG-relevanten Medikamenten verschrieben wurde. Haben nur wenige Praxen entsprechende Medikamente verschrieben, können die geschätzten PCG-Koeffizienten stark von der zufälligen Streuung beeinflusst sein.
  Als zweites stellt sich die Frage, wie die Wirkstoffmenge berücksichtigt werden soll. Im Risikoausgleich wird die Wirkstoffmenge in standardisierten Tagesdosen (DDD) ausgedrückt, um unterschiedliche Medikamente vergleichbar zu machen. Dies halten wir für ein sinnvolles Vorgehen. Die DDD-Menge pro Beobachtung kann dabei als kontinuierliche Variable spezifiziert werden, oder es können Gruppen gebildet werden, welche sich an der empirischen Verteilung der DDD-Mengen orientieren.

1.3.2 Empirischer Test der Morbiditätsindikatoren

Die Morbiditätsindikatoren Wahlfranchisen, Spital-im-Vorjahr und PCG werden auf der ersten Stufe in das Modell einbezogen (separat pro Facharztgruppe). Sie verbessern den statistischen Erklärungsgehalt des Gesamtmodells für die meisten Facharztgruppen deutlich. Wo sie statistisch nicht signifikant sind, haben sie zumindest das erwartete Vorzeichen (d. h. hohe Franchisen haben einen negativen Einfluss auf die Kosten, Spital-im-Vorjahr und PCG haben einen positiven Einfluss).

Bei der Indexberechnung hat sich zusätzlich gezeigt, dass nach der Korrektur um die AGG und die Morbiditätsvariablen die Praxen mit den höchsten durchschnittlichen Kosten nicht mehr «automatisch» als auffällig gelten. Sie sind nur dann auffällig, wenn sie hohe Kosten aufweisen, die von den erklärenden Variablen im Modell nicht erklärt werden können. Gestützt auf die empirischen Analysen und Erkenntnisse aus der internationalen Fachliteratur empfehlen wir, die Morbiditätsindikatoren Franchise, Spital-im-Vorjahr und PCG in das Modell aufzunehmen.

1.4 Indikatoren des Praxisstandortes

Auf der zweiten Stufe haben wir geprüft, ob neben dem Kanton auch noch weitere Charakteristika des Praxisstandorts berücksichtigt werden können. Konkret hatten wir die Hypothese aufgestellt, dass die Sozialhilfequote, die Einwohnerdichte oder der Ausländeranteil der Praxisge-

meinde den errechneten Praxiseffekt beeinflussen könnten. Diese Faktoren erwiesen sich jedoch in den ökonometrischen Schätzungen als statistisch nicht signifikant und auch von limitierter Grösse, so dass sie nicht zwingend ins Modell einbezogen werden müssen. Ein möglicher Grund für den limitierten Einfluss ist, dass diese Indikatoren nur auf Gemeindeebene zur Verfügung stehen und diese Grösse zu ungenau ist. So haben beispielsweise grosse Städte sowohl Quartiere mit sehr hohen als auch sehr geringen Sozialhilfequoten. Der Durchschnitt über die ganze Gemeinde ist möglicherweise wenig aussagekräftig.

1.5 Indexberechnung und Testgüte

Aus dem bereinigten Praxiseffekt wird anschliessend ein Index berechnet, welcher angibt, um wie viel Prozent die Kosten einer Praxis über dem erwarteten Wert liegen. Praxen, welche einen Indexwert von über 130 haben – also mehr als 30 Prozent über dem erwarteten Wert ihrer Fachärztgruppe liegen –, gelten als auffällig und werden weiter überprüft. Für eine definitive Beurteilung der Testgüte dieses Verfahrens müsste bekannt sein, welche Praxen effektiv unwirtschaftlich arbeiten. Diese Information ist aber leider nicht verfügbar, und es gibt auch kein Testverfahren, welches die Wirtschaftlichkeit von Arztpraxen erwiesenermassen wesentlich besser beurteilen könnte als das hier beschriebene statistische Screening (in der Fachsprache sagt man, es ist kein «Referenzstandard» verfügbar).

Um dennoch eine Aussage über die Treffergenauigkeit machen zu können, haben wir Simulationsrechnungen angewandt. Konkret haben wir den Praxen zufällig individuelle Praxiseffekte zugeordnet, von denen rund 10 Prozent die Grenze von 130 Prozent überschritten. Damit wissen wir, welche der simulierten Praxen ineffizient sind und welche nicht. Zusätzlich unterscheiden sich die simulierten Praxen in ihrem Patientenstamm und zufällig zugeordneten Kostenabweichungen. Die Zuordnung wiederholten wir über 100 Mal, um die Streuung der ermittelten Werte erfassen zu können.

Mit dieser Simulationsberechnung kann die Wahrscheinlichkeit beurteilt werden, dass eine Praxis wegen der zufälligen Streuung (z. B. wegen hohen Ausreisserwerten durch schwer erkrankte Patienten) falsch beurteilt wird. Die entsprechenden Kennzahlen sind in Tabelle 1 dargestellt. Besonders interessant ist bei der Wirtschaftlichkeitsprüfung der positive Vorhersagewert. Er gibt an, wie viel Prozent der als auffällig identifizierten Praxen (positiver Testwert) «richtig positiv» sind. Bei einer Indexberechnung mit dem Punktschätzer (ohne Berücksichtigung der Unsicherheit) sind dies über drei Viertel der positiv getesteten Praxen. Wird der Unsicherheitsindikator zur Indexberechnung verwendet, sind es über 96 Prozent. Durch das strengere Kriterium steigt jedoch auch die Wahrscheinlichkeit, dass eine richtig positive Praxis nicht mehr als solche erkannt wird. Dies schlägt sich in einer geringeren Sensitivität nieder: Anstatt knapp 90 Prozent mit dem Punktschätzer sind es nur noch gut 66 Prozent.

Bei der Interpretation der Kennzahlen muss beachtet werden, dass keine Praxisbesonderheiten (beispielsweise ein besonderes Leistungsspektrum) in der Simulation abgebildet wurden.

Tabelle 1 Beurteilung der Testgüte des Wirtschaftlichkeitsindexes

	Sensitivität	Spezifität	Positiver Vorhersagewert
Indexberechnung mit Punktschätzer	89.9%	96.9%	76.1%
Indexberechnung mit Untergrenze des Vertrauensbereichs	66.2%	99.8%	96.8%

Quelle: Eigene Darstellung, Polynomics.

1.6 Berechnungen mittels Individualdaten

1.6.1 Testgüte mittels Individualdaten

Drei Versicherer haben Abrechnungsdaten mit Bezug zu individuellen Patienten zur Verfügung gestellt (pseudonymisiert). Wir haben diese Daten einmal auf individueller Ebene und einmal aggregiert (nach der gleichen Methode wie der Sasis-Datensatz) ausgewertet. Gemäss unseren Analysen würden mit Individualdaten etwas mehr Praxen als auffällig identifiziert werden als mit aggregierten Daten. Konsistent dazu zeigte sich in der Simulation, dass bei Individualdaten die Anzahl falsch Negativer geringer war (höhere Sensitivität), es wird also ein grösserer Anteil der richtig Positiven entdeckt. Als Kehrseite zeigte sich jedoch, dass auch die Anzahl an falsch Positiven höher war (geringerer positiver Vorhersagewert).

Der Unsicherheitsindikator war bei der Berechnung mit Individualdaten systematisch geringer als bei einer Rechnung mit aggregierten Daten. Dies liegt daran, dass mit Individualdaten pro Praxis wesentlich mehr Datenpunkte zur Verfügung stehen. Es ist daher besser möglich, den systematischen Praxiseffekt von der zufälligen Schwankung zu trennen.

1.6.2 Einschätzung

Die Voraggregation führt immer zu einem Verlust an Information, die aggregierten Daten aus dem Daten- bzw. Tarifpool sind denn auch wesentlich weniger detailliert als patientenbezogene Abrechnungsdaten. In Bezug auf die vorliegende Frage haben unsere empirischen Analysen jedoch nicht ergeben, dass mit Individualdaten eine klare Verbesserung des Screenings erreicht werden kann. Die Anzahl falsch Positiver war in unseren Berechnungen mit aggregierten Daten sogar geringer.

Neu zu beurteilen wäre die Situation, wenn die Datenbasis um zusätzliche Informationen erweitert werden könnte. Diagnosebasierte Daten würden beispielsweise einen Vergleich der Kosten pro «Behandlungsepisode» erlauben, was sich in der US-amerikanischen Fachliteratur im Kontext des «physician profiling» durchgesetzt hat.

2 Einleitung

2.1 Ausgangslage

Das statistische Screening ist der erste Schritt der Wirtschaftlichkeitsprüfungen gemäss Art. 56 Abs. 6 KVG. Das Ziel des Screenings ist es, Leistungserbringer zu identifizieren, die ein auffälliges Kostenprofil aufweisen. Diese Leistungserbringer werden in weiteren Schritten vertieft abgeklärt. Um die weiteren Schritte möglichst effizient zu gestalten und bei den betroffenen Leistungserbringern keinen unnötigen Aufwand zu generieren, ist es wichtig, ein möglichst treffsicheres statistisches Verfahren einzusetzen.

Seit 2004 kommt für das statistische Screening die sogenannte ANOVA-Methode zum Einsatz (Roth und Stahel 2005). Grob gesprochen werden im ersten Schritt die logarithmierten mittleren Kosten pro Arzt um den Effekt der Alters- und Geschlechtsgruppe (im Folgenden: AGG) bereinigt. Auf einer zweiten Stufe folgt dann die Bereinigung um den Einfluss der Facharztgruppe und des Kantons.

Dieses Verfahren soll erweitert werden, einerseits durch den Einbezug zusätzlicher Morbiditätsindikatoren und andererseits – falls zweckmässig – auch durch Anpassungen an der Berechnungsmethodik. Zur Vorbereitung wurde im Frühjahr 2016 bei der Firma B,S,S. ein Gutachten in Auftrag gegeben, welches die bisher verwendete Methode formalisierte und Möglichkeiten zum Einbezug weiterer Morbiditätsindikatoren aufzeigte. Das Gutachten beinhaltet unter anderem auch Kriterien, nach denen die zusätzlichen Morbiditätsindikatoren ausgewählt werden können (Kaiser 2016).

2.2 Projektziele

Mit dem Gutachten Kaiser (2016) wurden die theoretischen Grundlagen zur Weiterentwicklung der statistischen Verfahren zur Wirtschaftlichkeitsprüfung geschaffen. Im vorliegenden Bericht werden diese datenbasiert umgesetzt und empirisch getestet. Dabei stehen folgende Fragen im Vordergrund:

- Diskussion des Schätzmodells im Lichte der internationalen Fachliteratur: Ist die durch Kaiser (2016) vorgeschlagene Spezifikation zweckmässig, beziehungsweise wie könnte die Spezifikation erweitert werden?
- Welche zusätzlichen Morbiditätsfaktoren eigenen sich für den Einbezug in das Modell? Wie können diese Morbiditätsfaktoren empirisch umgesetzt werden?
- Welche zusätzlichen Charakteristika der Arztpraxis, insbesondere der Praxisstandort, haben einen erheblichen Einfluss auf die Kosten pro Praxis? Wie können diese in das Modell eingefügt werden?
- Welche Treffgenauigkeit (Anzahl falsch positiv bzw. falsch negativ klassifizierte Praxen) wird mit dem statistischen Screening erreicht? Welche Erweiterungen des Schätzmodells könnten zu einer verbesserten Treffergenauigkeit führen?
- Könnte durch eine Berechnung mit patientenbezogenen Daten eine wesentliche Verbesserung der Treffergenauigkeit erreicht werden?

2.3 Aufbau des Berichtes

Die Arbeiten lassen sich grob in drei Teile gliedern: einen theoretisch-konzeptionellen Teil, einen empirischen Teil mit den aggregierten Daten der Sasis AG und einen empirischen Teil mit den Individualdaten. Der konzeptionelle Teil beginnt in Kapitel 3 mit einem Überblick über die wichtigsten Erkenntnisse aus der Fachliteratur. In Kapitel 4 folgt eine theoretische Diskussion der durch Kaiser (2016) empfohlenen, zweistufigen Methode und möglichen Erweiterungen. Zudem wird eine Auswahl an Morbiditätsindikatoren getroffen, deren Einbezug danach empirisch getestet wird.

In den Kapiteln 5 bis 8 folgen Auswertungen anhand der Daten der Sasis AG. In Kapitel 5 beschreiben wir die Datenaufbereitung, insbesondere die Definition der Zielvariablen und der Morbiditätsindikatoren. Diese Variablen werden in Kapitel 6 genutzt, um die erste Stufe des ökonometrischen Schätzmodells zu berechnen. Die zweite Stufe und die Indexberechnung beschreiben wir in Kapitel 7. In Kapitel 8 zeigen wir eine Simulationsrechnung, welche dazu dient, die Treffgenauigkeit des statistischen Screeningverfahrens zu beurteilen.

In Kapitel 9 folgen schliesslich die Berechnungen auf Basis der Versichererdaten (inkl. Zuordnung zu einzelnen Patienten). Dabei testen wir zuerst unterschiedliche Varianten der Indexberechnung, anschliessend erfolgt eine Simulationsrechnung mit dem Ziel, die Testgüte der aggregierten Daten mit Individualdaten vergleichen zu können. Der Bericht schliesst mit einem Fazit in Kapitel 10.

Im Anhang (Kapitel 11) finden sich Detailauswertungen, ökonometrische Vertiefungen und weitergehende Beschreibungen technischer Natur. Der Anhang richtet sich vor allem an interessierte Leser mit den entsprechenden statistischen Kenntnissen und soll den Hauptteil von unnötigen Detailinformationen entlasten.

Wegen der hohen Komplexität des Themas werden auch im Hauptbericht statistische und ökonometrische Fachbegriffe verwendet. Dies ist der wissenschaftlichen Genauigkeit und der Transparenz geschuldet, die wir in diesem Bericht anstreben. Die statistischen und ökonometrischen Überlegungen und Umsetzungen sollen möglichst vollständig aufgeführt und begründet sein, damit sie sauber nachvollzogen und überprüft werden können. Sowohl in der Zusammenfassung (Kapitel 1) als auch im Fazit (Kapitel 10) haben wir versucht, die wichtigsten Erkenntnisse möglichst ohne statistische Fachbegriffe darzulegen.

3 Literaturüberblick

Die statistische Beurteilung der Wirtschaftlichkeit von Ärzten wird in der Fachliteratur unter dem Stichwort «Physician Profiling» zusammengefasst. Wichtige Studien kommen dabei aus den USA und – zu einem kleineren Teil – aus den Niederlanden, also aus Ländern, welche (ebenso wie die Schweiz) ein eher wettbewerbliches Gesundheitswesen haben. In Ländern, wo das Gesundheitswesen stärker staatlich gesteuert ist, kommen eher Budgets zum Einsatz. Bei der Berechnung der Budgets stellen sich zum Teil die gleichen methodischen Probleme wie bei der Betrachtung der Wirtschaftlichkeit, insbesondere muss ebenfalls eine Erfassung der Morbidität des Patientenstamms stattfinden.

Für die Studie haben wir eine zielgerichtete Literaturrecherche durchgeführt, welche auf die Beantwortung der folgenden Fragen ausgerichtet war:

- Welche Indikatoren/Methoden werden zur Risikoadjustierung eingesetzt?
- Welche Methoden werden zur Morbiditätskorrektur und zur Berechnung von Indices («Physician Scores») eingesetzt?
- Wie wird die Eignung/Zuverlässigkeit des Verfahrens evaluiert?

Folgende Stichwörter haben wir bei der Literatursuche verwendet: «Physician Profiling», «Health care cost metrics» «Risk adjustment», «Capitation». Aus der Resultatmenge wurden Studien ausgewählt, welche empirische Auswertungen mit Abrechnungsdaten enthalten («real data studies») und auf ambulante Arztpraxen ausgerichtet sind.

Das Kapitel ist thematisch aufgebaut nach den oben erwähnten drei Hauptfragen. Es werden also keine einzelnen Arbeiten integral zusammengefasst. Die wichtigsten Quellen, welche von drei Forschergruppen stammen, sollen hier trotzdem aufgelistet werden:

- **Rand Cooperation (Adams 2009; Adams, Mehrotra und McGlynn 2010; Adams et al. 2010)**
  Mehrjährige Forschungsarbeit zur Beurteilung der in den USA gängigen Methoden des Physician Profiling; Entwicklung eines Indikator der Zuverlässigkeit
- **University of Michigan (Thomas et al., 2004a, 2004b; Thomas und Ward, 2006)**
  Mehrjährige Forschungsarbeit zum Vergleich verschiedener Methoden des Physician Profiling; Beurteilung der Treffgenauigkeit mittels Simulation
- **Universität Rotterdam (Eijkenaar und van Vliet 2014; Eijkenaar und van Vliet 2013)**
  Mehrjährige Forschungsarbeit zur Beurteilung unterschiedlicher statistischer Methoden zur Indexberechnung

3.1 Morbiditätskorrektur

3.1.1 Morbiditätsinformation

Diagnostische Informationen sind die mit Abstand wichtigste und am häufigsten eingesetzte Morbiditätsinformation in der Fachliteratur. Die detaillierten Diagnosen werden nach einem einheitlichen Codierungssystem wie ICD-9 oder ICD-10 erfasst und anschliessend durch ein softwaregestütztes Klassifikationssystem (Grouper) zu einer überschaubaren Anzahl an Analysegruppen zusammengefasst (Thomas et al., 2004a; Adams et al., 2010a; Wasem et al., 2010; Kristensen et al., 2014).

Stehen keine diagnostischen Informationen zur Verfügung, werden Morbiditätsindikatoren aus der beobachteten Leistungsinanspruchnahme gebildet. Besonders bewährt haben sich hier pharmazeutischen Kostengruppen (PCG) (Eijkenaar und van Vliet, 2013; von Rotz et al., 2008). Das Klassifikationssystem (Grouper) wird entwickelt, indem Medikamente anhand ihrer Hauptwirkstoffe zu Indikationsgebieten zugeteilt werden. Patienten, welche Medikamente aus einem spezifischen Indikationsgebiet bezogen haben, werden in die entsprechenden PCG eingeteilt.

Ein alternativer Indikator des Gesundheitszustands sind die individuellen Vorjahreskosten (von Rotz, Kunze und Beck 2008). Diese lassen in der Regel eine sehr gute Prognose der Kosten im Folgejahr zu. Sie werden jedoch dafür kritisiert, dass sie nicht unabhängig vom Arztverhalten sind (keine exogene Variable, Van de Ven und Ellis 2000).

Weitere Indikatoren, welche mit den Kosten pro Patient korreliert sind, sind Charakteristika der Wohnregion wie die Urbanisierung, der Anteil an Ausländern, oder sozioökonomische Indikatoren (Eijkenaar und van Vliet 2013).

3.1.2 Risikoadjustierung: Modellwahl

Episodes of Care

Besonders in der US-amerikanischen Literatur ab Anfang der 2000er Jahre wird meistens mit dem Episodes-of-Care-Ansatz gearbeitet (Adams, Mehrotra und McGlynn 2010; Thomas, Grazier und Ward 2004a; Thomas, Grazier und Ward 2004b). Durch spezialisierte Softwaretools werden alle Abrechnungen (bspw. Arztleistungen, Labor, Medikamente etc.) bestimmten Diagnosen und Zeiträumen zugeordnet. Soweit möglich werden die Episoden den Ärzten zugeteilt. Dies geschieht nach festen Zuordnungsregeln. Beispiele von solchen Zuordnungsregeln können sein:

1. Bei jedem Arzt, der mehr als 20 Prozent der Arztleistungen der Episode erbracht hat, zählt diese.
2. Die Episode zählt beim Arzt mit dem höchsten Anteil Arztkosten, solange es über 30 Prozent sind.

Im ersten Fall erlaubt die Regel, dass eine Episode bei mehreren Ärzten in die Beurteilung (Physician Score) eingeht. In beiden Fällen ist es möglich, dass eine Episode überhaupt nicht zugeteilt wird, weil kein Leistungserbringer als der «entscheidende» Verursacher identifiziert werden kann.

Bei einem episodenbasierten Ansatz werden viele sehr kleinteilige Gruppen gebildet, die sich gegenseitig ausschliessen. Die durchschnittlichen Kosten innerhalb der gleichen Episode gelten als direkt vergleichbar, so dass keine zusätzliche Morbiditätskorrektur vorgenommen wird.

Multivariate Regressionsmodelle

Multivariate Regressionsmodelle sind darauf ausgerichtet, den durchschnittlichen Einfluss von erklärenden Faktoren auf eine Zielvariable (z. B. die Kosten) zu berechnen. Mit dieser Methode kann abgeschätzt werden, welche Auswirkungen bestimmte Morbiditätsvariablen (z. B. Alter, Geschlecht, pharmazeutische Kostengruppen) auf die erwarteten Kosten einer Arztpraxis haben. Die Beurteilung des Arztes erfolgt dann meist durch die Gegenüberstellung der erwarteten Kosten eines Arztes mit seinen beobachteten Kosten. In der Beurteilung von Arztpraxen wurden

Regressionsmodelle beispielsweise eingesetzt durch von Rotz et al. (2008), Eijkenaar und van Vliet (2013) sowie Wasem et al. (2010).

Eijkenaar und van Vliet (2014) gehen detailliert auf unterschiedliche Aspekte der Modellwahl ein. Sie vergleichen eine Vielzahl an Regressionsmethoden, welche in der gesundheitsökonomischen Literatur häufig verwendet werden, unter anderem Ordinary Least Squares (OLS), Generalized Linear Models (GLM), Two-part Models oder Multilevel-Modelle («Random Effects»). Letztere Modelle erlauben es, dafür zu kontrollieren, dass pro Arzt mehrere Beobachtungen zur Verfügung stehen. Die empirische Untersuchung ergab, dass die Resultate im Allgemeinen nicht sehr stark voneinander abweichen. Die Autoren folgern (Zitat Eijkenaar und van Vliet, 2014):

«Differences were relatively small and the choice of model may not be as important as other choices such as the set of risk-adjusters, definition of performance index, and method for categorizing provider performance.»

Die Autoren geben den Modellen auf der Basis von OLS einen leichten Vorzug, weil hier die gleiche Spezifikation für die Modellierung von unterschiedlichen Zielvariablen verwendet werden kann und die Modellgüte jeweils nur leicht schlechter war als bei spezifischeren Modellen.

3.2 Indexberechnung

Die Resultate der Morbiditätsbereinigung müssen anschliessend in eine Beurteilung von einzelnen Praxen übergeführt werden (Indexberechnung). Der in der Literatur am häufigsten verwendete Ansatz zur Indexberechnung ist die «Predictive Ratio» (Thomas, Grazier und Ward 2004a; von Rotz, Kunze und Beck 2008; Wasem, Lux und Dahl 2010). Dazu werden die beobachteten Kosten pro Praxis (*i*) den erwarteten Kosten gegenübergestellt. Die «Predictive Ratio» ist dementsprechend grösser als eins, wenn die beobachteten Kosten über den erwarteten liegen und kleiner als eins, wenn sie darunterliegen.

$$\text{Predictive Ratio}_i = \frac{\text{Mittelwert beobachtete Kosten}_i}{\text{Mittelwert erwartete Kosten}_i}$$

Es gibt dabei unterschiedliche Varianten, wie der «Mittelwert erwartete Kosten» berechnet werden kann. Bei einem Ansatz mit Behandlungsepisoden werden zuerst die durchschnittlichen Kosten pro Episode über alle Ärzte berechnet. Die Durchschnitte werden gewichtet mit der Häufigkeitsverteilung desjenigen Arztes, welcher beurteilt werden soll. Bei einem Regressionsansatz können die erwarteten Kosten pro Patient direkt berechnet werden. Es wird der Mittelwert über alle Patienten des zu beurteilenden Arztes genommen. Dieser Wert entspricht den fiktiven Kosten dieser Patienten, wenn sie von einem durchschnittlichen Arzt behandelt würden.

Als Alternative zur Predictive Ratio berechnen Thomas et al. (2004b) einen Index mittels der «Standardized Cost Difference» (SCD). Hier wird die absolute Differenz zwischen dem Mittelwert der beobachteten und erwarteten Kosten als Indikator verwendet. Er wird mit der Standardabweichung der erwarteten Kosten (σ) über alle Praxen gewichtet, geteilt durch die Quadratwurzel der Anzahl Patienten (N) der Praxis i.

$$\text{SCD}_i = \frac{\text{Mittelwert beobachtete Kosten}_i - \text{Mittelwert erwartete Kosten}_i}{\sigma/\sqrt{N_i}}$$

Der grosse Vorteil der SCD ist, dass die Anzahl Patienten direkt berücksichtigt wird. Bei einer geringen Anzahl an Patienten kann der Index nicht mit der gleichen Sicherheit berechnet werden wie bei grossen Praxen. Es ist daher bei kleinen Praxen häufiger, dass sehr hohe oder sehr

niedrige Indexwerte auftreten (Thomas, Grazier und Ward 2004b). Nachteilhaft an der SCD ist jedoch, dass kein Indexwert resultiert, welcher als Zahl interpretierbar ist (z. B. in Prozent). Die Praxen können nach der SCD lediglich rangiert werden.

3.3 Evaluation des Profilings

Grundsätzlich kann die Güte eines Testverfahrens dann am besten beurteilt werden, wenn ein anderes Testverfahren zur Verfügung steht, von dem bekannt ist, dass es eine sehr hohe Treffsicherheit aufweist. In der Fachliteratur bezeichnet man dies als «Referenzstandard». Bei der Beurteilung der Wirtschaftlichkeit ist jedoch leider kein Referenzstandard verfügbar, denn es existiert kein Verfahren, welches exakt und ohne jeden Zweifel feststellen könnte, welche Praxen wirtschaftlich arbeiten.

In den zitierten Studien wurden zwei Verfahren angewandt, um die Güte des Profilings dennoch zu beurteilen. Der Indikator der «Zuverlässigkeit» wurde von Adams et al. (2010a) entwickelt und später von Eijkenaar und van Vliet (2013) verwendet. Der Indikator stützt darauf ab, dass pro Praxis mehrere Beobachtungen zur Verfügung stehen. Er vergleicht die Varianz (σ^2) der Beobachtungen *innerhalb* einer Praxis mit der Varianz der Beobachtungen *zwischen* den Praxen. Der Einfluss einer Praxis kann dann am zuverlässigsten bestimmt werden, wenn die Streuung zwischen den Praxen im Vergleich zur Streuung innerhalb der Praxis gross ist.

$$\text{Zuverlässigkeit} = \frac{\sigma^2_{\text{zwischen den Praxen}}}{\sigma^2_{\text{zwischen den Praxen}} + \sigma^2_{\text{innerhalb einer Praxis}}}$$

Die Autoren kommen zum Schluss, dass die Varianz innerhalb einer Praxis vergleichsweise bedeutend ist. Das Risiko von Fehlklassifikationen ist deshalb nicht vernachlässigbar.

Thomas und Ward (2006) nutzen eine Simulation zur Beurteilung der Güte des Profilings. Auch sie nutzen die Tatsache, dass pro Praxis mehrere Beobachtungen (Episoden) zur Verfügung stehen. In einem ersten Schritt teilen sie die Arztpraxen anhand ihrer berechneten SCD in drei Gruppen ein (effizient, mittel, ineffizient). Danach werden drei Datensätze von einzelnen Episoden gebildet. Datensatz eins enthält die Episoden aller als «effizient» klassifizierten Ärzte, Datensatz zwei alle Episoden der «mittleren Ärzte» und Datensatz drei alle Episoden der als «ineffizient» klassifizierten Ärzte. Zudem wird berechnet, welche Art von Episoden für welche Facharztgruppe wie häufig vorkommen.

Auf dieser Grundlage werden fiktive Praxen simuliert. Nehmen wir beispielsweise an, es solle eine «effiziente» Praxis für Allgemeinmedizin simuliert werden und die durchschnittliche Praxis für Allgemeinmedizin behandle 300 Episoden «Diabetes» und 700 Episoden «akute Sinusitis».[1] Es würde dann aus dem Datensatz der Episoden der effizienten Ärzte (Datensatz eins) zufällig gezogen, und zwar 300 Diabetes-Episoden und 700 Sinusitis-Episoden. Diese Episoden bilden zusammen eine fiktive Praxis.

Für die simulierten Praxen wird dann ein «Profiling» erstellt. Alle fiktiven Praxen, die aus den Episoden der effizienten Ärzte gezogen wurden, sollten auch als effizient beurteilt werden. So kann die Sensitivität, Spezifität und der Predictive Error des Testes berechnet werden. Bei den Ärzten der Allgemeinmedizin («Family Practice») lag der Predictive Error bei der Entdeckung von ineffizienten Praxen bei rund 20 Prozent. Rund ein Fünftel der als ineffizient beurteilten Praxen stammte also nicht aus der entsprechenden Kategorie.

[1] Die Einschränkung auf zwei Diagnosen ist lediglich beispielhaft.

4 Theorie

4.1 Schätzmodell

Entsprechend dem Projektauftrag bildet der im Gutachten von Kaiser (2016) präsentierte Modellrahmen den Ausgangspunkt für unsere Analysen. Im Folgenden diskutieren wir Vor- und Nachteile des vorgeschlagenen zweistufigen Modells und gehen auf ausgewählte Aspekte des Rechenverfahrens ein.

4.1.1 Erste Stufe: Fixed-Effects-Modell zur Berechnung von praxisspezifischen Effekten

Das Kernstück der Wirtschaftlichkeitsprüfung bildet das Fixed-Effects-Modell in Gleichung (1). Das Modell bestimmt die logarithmierten Kosten y_{ij} pro Praxis i und Alters- und Geschlechtsgruppe (AGG) j in Abhängigkeit eines Effekts pro AGG, den Morbiditätsvariablen X_{ij} und einer praxisspezifischen Konstanten (a_i),

$$\ln y_{ij} = \mathrm{AGG}_j\beta_1 + \mathrm{X}_{ij}\,\beta_2 + a_i + \varepsilon_{ij} \quad \text{für } i \in \{FAG_f\} \tag{1}$$

Die Schätzung der praxisspezifischen Konstanten a_i wird dadurch ermöglicht, dass pro Arztpraxis mehrere Beobachtungen zur Verfügung stehen (Panelstruktur). Zur weiteren Verwendung wird die Konstante um die mittlere Konstante pro Facharztgruppe bereinigt. Die resultierende Differenz (praxisspezifischer Effekt) gibt an, ob die durchschnittlichen Kosten einer Praxis gegebenen die erklärenden Variablen höher oder niedriger als der Gesamtdurchschnitt sind.

Die Variable ε_{ij} bezeichnet den stochastischen Störterm, welcher durch das Modell nicht erklärbare, unsystematische Kostenunterschiede auffängt. Da nicht davon auszugehen ist, dass die erklärenden Variablen wie beispielsweise die Altersstruktur des Patientenstamms bei allen Facharztgruppen den gleichen Einfluss auf die Kosten haben, wird das Modell (1) pro Facharztgruppe (FAG) separat geschätzt. Die Schätzung wird gewichtet mit der Anzahl verfügbarer Beobachtungen pro AGG und Praxis.

Vorteile des Fixed-Effects-Modells

- **Vorteile gegenüber dem Mittelwertvergleich und der linearen Regression: Die zufällige Schwankung wird vom unerklärten Praxiseffekt getrennt**

  In diesem Abschnitt diskutieren wir die Vorteile eines Fixed-Effects-Modells gegenüber einem reinen Mittelwertvergleich, und einer Modellierung der Praxiskosten mit einer gewöhnlichen linearen Regression.
  Bei einem Mittelwertvergleich können keine Einflussfaktoren berücksichtigt werden. Stattdessen werden die durchschnittlichen Kosten einer Praxis direkt mit dem Gesamtdurchschnitt verglichen. Diese Vorgehensweise geht implizit davon aus, dass alle Praxen einen homogenen Patientenstamm haben: Jede Abweichung vom Mittelwert wird als Über- oder Untereffizienz betrachtet.
  Mit einer linearen Regression werden bei der Bestimmung der erwarteten Kosten erklärbare Kostenunterschiede berücksichtigt, indem entsprechende Variablen (wie bspw. die AGG oder PCG) in die Schätzgleichung aufgenommen werden (siehe z. B. von Rotz et al., 2008). Diese Vorgehensweise lässt somit Heterogenität unter den Praxen zu. Sie betrachtet jedoch weiterhin sämtliche Abweichungen von den so ermittelten erwarteten Kosten als Unter- oder Übereffizienz.

Das Fixed-Effects-Modell erweitert die gewöhnliche lineare Regression, indem bei der Schätzung die nicht durch das Modell erklärten Unterschiede in zwei Komponenten aufgeteilt werden: erstens eine systematische Komponente pro Praxis (a_i in Gleichung (1)), zweitens einen zufälligen Term (ε_{ij} in Gleichung (1)). Für die Beurteilung der Praxis wird nur die systematische Komponente (a_i) verwendet. Diese Komponente erfasst dabei nicht nur alle beobachtbaren, sondern auch alle nicht beobachtbaren Faktoren, welche die systematischen Unterschiede in den praxisspezifischen Kosten erklären können. Anders ausgedrückt handelt es sich um Abweichungen von den erwarteten Kosten, welche spezifisch für eine Praxis sind.
Der zufällige Term kann Messfehler oder zufällige Abweichungen aufgrund des individuellen Leistungsbedarfs beinhalten. Er wird nicht zur Beurteilung der Praxis verwendet.

- **Vorteil gegenüber dem Random-Effects-Modell: Erklärende Variablen dürfen mit dem Praxiseffekt korreliert sein**

  Stehen mehrere Beobachtungen pro Praxis zur Verfügung (Panelmodell), kann der praxisspezifische Effekt grundsätzlich auf zwei verschiedene Arten modelliert werden: Mittels Random-Effects-Modell oder mittels Fixed-Effects-Modell. Für Details zu diesen beiden Modellvarianten siehe beispielsweise Wooldridge (2010). Die Fixed-Effects-Variante hat insbesondere den Vorteil, dass sie Korrelationen zwischen dem praxisspezifischen Effekt a_i und den anderen erklärenden Variablen zulässt. Es ist sehr wahrscheinlich, dass es in unserem Fall unbeobachtete Praxischarakteristika gibt, welche Einfluss auf den Patientenstamm und damit auf die erklärenden Variablen in Modell (1) haben. Wir müssen somit davon ausgehen, dass Korrelationen zwischen der praxisspezifischen Konstante und den anderen erklärenden Variablen vorliegen. Damit bietet sich die Fixed-Effects-Spezifikation an.

Nachteile des Fixed-Effects-Modells

- **Praxischarakteristika können nicht direkt auf der ersten Stufe verwendet werden**

  In einem Fixed-Effects-Modell wie es durch die Gleichung (1) definiert ist, ist es nicht möglich, Variablen zu berücksichtigen, welche pro Praxis nicht variieren. Der Einfluss solcher Variablen könnte rechnerisch nicht vom Einfluss des praxisspezifischen Effekts getrennt werden. Eine Berücksichtigung ist nur mit einer Korrektur auf einer zweiten Stufe möglich. Diese Korrektur behandeln wir in Abschnitt 4.1.2.

- **Alle nicht beobachteten Praxisbesonderheiten gehen in den Fixed Effect ein**

  Als Nachteil einer Fixed-Effects-Spezifikation in der Effizienzmessung ist zu sehen, dass *alle* nicht beobachteten Praxisbesonderheiten in den praxisspezifischen Effekt eingehen (Schmidt und Sickles 1984). Hat beispielsweise ein Arzt ein bestimmtes Leistungsspektrum, welches mit hohen Kosten verbunden ist, wird ihm ein hoher spezifischer Praxiseffekt zugerechnet. In diesem Fall ist der hohe Praxiseffekt jedoch nicht auf Ineffizienz, sondern auf nicht beobachtete Praxisbesonderheiten zurückzuführen. Dies muss bei der Interpretation der Ergebnisse zwingend beachtet werden.
  In unserem Kontext fällt dieser Aspekt der Fixed-Effects-Schätzung nicht so schwer ins Gewicht, da das statistische Screening dazu verwendet wird, Praxen mit auffälligen Kostenprofilen zu identifizieren. Diese werden anschliessend eingehend überprüft. Praxisbesonderheiten wie beispielsweise ein spezifisches Leistungsspektrum, welche zu hohen Kosten führen, müssten in der Praxisprüfung analysiert werden.

Fazit

Für das gegenwertige Problem überwiegen aus unserer Sicht die Vorteile des Fixed-Effects-Modells. Es zeichnet sich besonders dadurch aus, den praxisspezifischen Effekt vom Einfluss der erklärenden Variablen und vom zufälligen Störterm zu trennen.

4.1.2 Zweite Stufe: Bereinigung um praxisspezifische Variablen

Mit der Fixed-Effects-Spezifikation in Modell (1) können keine Variablen berücksichtigt werden, welche auf Praxisebene konstant sind. Dazu ist ein zweistufiges Vorgehen nötig (Kaiser 2016). Nach der Fixed-Effects-Regression gemäss Modell (1) wird in der zweiten Stufe eine lineare Regression mit dem in der ersten Stufe ermittelten spezifischen Praxiseffekt $\hat{a}_i$ als abhängige Variable durchgeführt. Als erklärende Variablen werden der Kanton KT_i, die Facharztgruppe FAG_i und weitere Faktoren Z_i berücksichtigt, ergänzt um eine Schätzkonstante δ_0 und einen Störterm u_i.

$$\hat{a}_i = \delta_0 + \delta_1 KT_i + \delta_2 Z_i + \delta_3 FAG_i + u_i \quad (2)$$

Die Facharztgruppen FAG_i können als facharztgruppenspezifischer Fixed Effect interpretiert werden und sorgen zusätzlich dafür, dass der Durchschnitt über die Residuen pro Facharztgruppe 0 beziehungsweise im Logmodell gleich 1 ist. Gleichung (2) kann zu

$$\hat{u}_i = \hat{a}_i - \hat{\delta}_0 - \hat{\delta}_1 KT_i - \hat{\delta}_2 Z_i - \hat{\delta}_3 FAG_i \quad (3)$$

umgeschrieben werden. Damit wird deutlich, dass das geschätzte Residuum $\hat{u}_i$ dem spezifischen Praxiseffekt $\hat{a}_i$ aus der ersten Stufe, korrigiert um die erklärenden Variablen aus der zweiten Stufe entspricht.

Vorteil des zweistufigen Vorgehens

Der Vorteil dieses Vorgehens ist, dass auf der ersten Stufe ein Fixed-Effects-Modell verwendet werden kann (mit den im letzten Abschnitt genannten Vorteilen) und trotzdem Variablen einbezogen werden können, welche pro Arztpraxis konstant sind. In der Fachliteratur werden Ansätze diskutiert, wie sich dasselbe mit einer einstufigen Schätzung erreichen lässt. Es hat sich bislang jedoch noch kein Standard herausgebildet (siehe z. B. Plümper und Troeger, 2007; Beck, 2011).

Nachteil des zweistufigen Vorgehens

Die zweite Stufe kann als Hilfskonstrukt angesehen werden, weil die Fixed-Effects-Methode nicht in der Lage ist, Variablen zu berücksichtigen, welche auf der Praxisebene konstant sind. In der Literatur werden solche Schätzungen häufig verwendet (siehe z. B. Kristensen et al., 2014). Mit dem zweistufigen Vorgehen ist es jedoch nicht möglich, die Kovarianz der Störterme der beiden Stufen zu berechnen. Geht man beispielsweise davon aus, dass die Störterme der ersten Schätzung in gewissen Kantonen stärker variieren als in anderen, könnte die Schätzung der Standardfehler für die Kantonsvariable auf der zweiten Stufe verzerrt sein.

Fazit

Aus unserer Sicht ist die zweite Stufe wie von Kaiser (2016) beschrieben ein valides Vorgehen, um den berechneten praxisspezifischen Effekt um Praxischarakteristika wie beispielsweise den Praxisstandort zu bereinigen.

4.2 Indexberechnung

In dem von Kaiser (2016) vorgeschlagenen Modell werden die Kosten in logarithmierter Form verwendet (vgl. Gleichung 1). Diese Transformation reduziert die Rechtsschiefe der Verteilung der Kosten (Zielvariable) und der Residuen. Die Logarithmierung der Kosten hat zusätzlich den Vorteil, dass der spezifische Praxiseffekt ($\hat{a}_i$) sowie das Residuum der zweiten Stufe ($\hat{u}_i$) als relative Grössen interpretiert werden können. Approximativ kann der prozentuale Praxiseffekt mit dem Index $100 \cdot \exp(\hat{u}_i)$ berechnet werden.

Die Logarithmierung hat hingegen den Nachteil, dass der Mittelwert von $exp(\hat{u}_i) \neq 1$ ist und damit der Mittelwert des Index nicht automatisch 100 beträgt (siehe Kaiser, 2016, S. 8). Da jede Praxis relativ zum Mittelwert ihrer Vergleichsgruppe beurteilt werden sollte und dieser Mittelwert als erwartete Kosten zu verstehen ist, sollte eine Praxis mit Kosten in der Höhe dieses Mittelwerts einen Index von 100 aufweisen.

Deshalb werden die Indexwerte mit einem Skalierungsfaktor normiert. Die Normierung erfolgt pro Facharztgruppe f indem $\exp(\hat{u}_i)$ mit $S_f = \left(\frac{1}{\frac{1}{N}\sum_i^N \exp(\hat{u}_i)}\right) * 100$ multipliziert wird. Dadurch ergibt sich für den Mittelwert über alle Indexwerte

$$\widehat{U}_i = S_f \times \exp(\hat{u}_i) \text{ für alle } i \text{ und } f \in \{1,2 \dots, F\} \quad (4)$$

ein Wert von 100 pro Facharztgruppe.

4.3 Gewichtung

In dem von Kaiser (2016) vorgeschlagenen Verfahren wird die erste Stufe gewichtet mit der Anzahl Patienten pro Arzt und AGG. Dies hat den wünschenswerten Effekt, dass Beobachtungen, die auf einer grösseren Anzahl Patienten basieren, einen stärkeren Einfluss ausüben. Die Schätzung wird präziser und effizienter. Würde man keine erklärenden Variablen einbeziehen, entspräche das gewichtet geschätzte Fixed-Effects-Modell einem Mittelwertvergleich der Kosten pro Patient und Arzt. Bezüglich der Berechnungsart findet so keine Abkehr von dem bisherigen Verfahren statt. Die Morbidität des Patientenstamms kann aber besser berücksichtigt werden.

Zusätzlich zu der von Kaiser (2016) vorgeschlagenen Gewichtung auf der ersten Stufe empfehlen wir, auch die Regression auf der zweiten Stufe und die Berechnung des Skalierungsfaktors zu gewichten, diesmal mit der Anzahl Patienten pro Praxis. Die Gründe sind die gleichen wie im vorigen Absatz erwähnt.

4.4 Unsicherheitsindikator und Berechnung einer Untergrenze des Indexes

4.4.1 Unsicherheitsindikator für den praxisspezifischen Effekt

Um eine Aussage über die Präzision der Schätzung der praxisspezifischen Effekte $\hat{a}_i$ auf der erste Stufe zu treffen, wird ein Unsicherheitsindikator (ähnlich einem Standardfehler) berechnet. Um die Berechnung nachzuvollziehen, rufen wir uns zuerst die Berechnung des praxisspezifischen Effektes ($\hat{a}_i$) in Erinnerung. Wie Kaiser (2016, S. 6) ausführt, kann der praxisspezifische Effekt als durchschnittliches Residuum pro Praxis berechnet werden ($\hat{a}_i + \hat{\varepsilon}_{ij} = \widehat{\omega}_{ij}$). Der betreffende Schätzer pro Praxis i ist in Gleichung (5) beschrieben, wobei J_i die Anzahl Be-

obachtungen (Anzahl AGG) pro Praxis bezeichnet, N_{ij} die Anzahl Patienten pro Praxis und AGG und N_i die Anzahl Patienten pro Praxis.

$$\hat{a}_i = \sum_{j=1}^{J_i} \frac{N_{ij}}{N_i}\left(y_{ij} - \hat{y}_{ij}\right) = \sum_{j=1}^{J_i} \frac{N_{ij}}{N_i}\left(y_{ij} - \left(\mathrm{AGG}_j\hat{\beta}_1 + \mathrm{X}_{ij}\hat{\beta}_2\right)\right)$$
$$= \sum_{j=1}^{J_i} \frac{N_{ij}}{N_i}\ \hat{\omega}_{ij} \quad \text{mit } \hat{\omega}_{ij} = \hat{a}_i + \hat{\varepsilon}_{ij} \text{ und } \sum_{j=1}^{J_i} \frac{N_{ij}}{N_i} = 1 \tag{5}$$

Wir schlagen vor, bei der Berechnung des Vertrauensindikators ein analoges Rechenverfahren anzuwenden. Es basiert auf den Residuen und Anzahl Beobachtungen pro Praxis. Zur Bildung des Unsicherheitsindikators verwenden wir die Formel zur Berechnung der Varianz des Mittelwerts einer Zufallsvariablen x mit einer beliebigen Wahrscheinlichkeitsverteilung und tatsächlichem Mittelwert μ sowie Varianz σ_x^2. Bei einer Stichprobe mit n unabhängigen Beobachtungen entspricht der Schätzer für den Mittelwert $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i$ und die Varianz des Schätzers entspricht

$$Var(\hat{\mu}) = Var\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} Var(x_i) = \frac{1}{n^2}\, n\, \sigma_x^2 = \frac{1}{n}\, \sigma_x^2 \tag{6}$$

Für die unbekannte Varianz σ_x^2 kann der Punktschätzer $\hat{\sigma}_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \hat{\mu})^2$ verwendet werden (Schira 2009). In Modell (1) entspricht der praxisspezifische Effekt $\hat{a}_i$ dem Mittelwert der Residuen $\hat{\omega}_{ij}$ pro Arzt, siehe Gleichung (5). Der Schätzer für die Varianz der Residuen pro Praxis ist demnach:

$$\hat{\sigma}^2_{\hat{\omega}_{ij}} = \frac{J_i}{J_i - 1}\sum_{j=1}^{J_i} \frac{N_{ij}}{N_i}\left(\hat{\omega}_{ij} - \hat{a}_i\right)^2. \tag{7}$$

Wir definieren den Unsicherheitsindikator wie folgt:

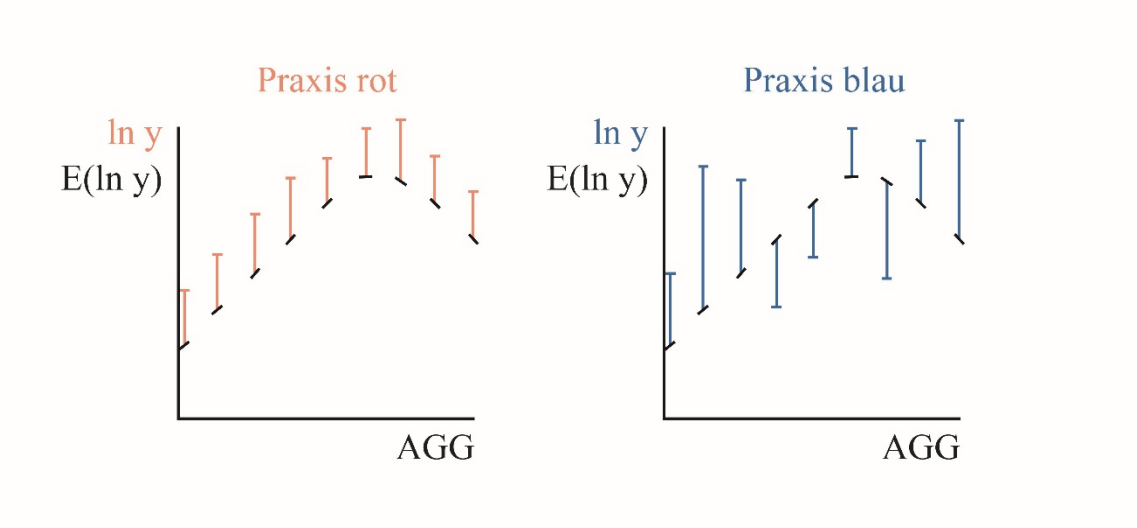
$$\text{Unsicherheitsindikator}_{\hat{a}_i} = \sqrt{\frac{1}{J_i}\,\hat{\sigma}^2_{\hat{\omega}_{ij}}} = \sqrt{\frac{1}{J_i}\,\frac{J_i}{J_i - 1}\sum_{j=1}^{J_i} \frac{N_{ij}}{N_i}\left(\hat{\omega}_{ij} - \hat{a}_i\right)^2} \tag{8}$$

Der Unsicherheitsindikator soll die Unsicherheit abbilden, mit welcher der praxisspezifische Effekt geschätzt werden kann. Intuitiv lässt sich der Indikator wie folgt interpretieren: Weicht eine Praxis in allen AGG im ähnlichen Umfang von den durch das Modell vorhergesagten Kosten ab, ist der Unsicherheitsfaktor gering. Der praxisspezifische Effekt ($\hat{a}_i$) beschreibt die Abweichung in den einzelnen AGG sehr gut. Weicht die Praxis aber beispielsweise in einer (oder einigen) AGG stark positiv ab, in anderen jedoch kaum oder sogar negativ, ist der Unsicherheitsindikator hoch.

In Abbildung 2 wird dies anhand zweier fiktiver Beispielspraxen illustriert: Die schwarzen Punkte stellen die Regressionsgerade dar (= erwarteten Werte einer Praxis bei gegebenen Morbiditätsvariablen). Die senkrechten farbigen Linien sind die Abweichungen der beobachteten Werte vom erwarteten Wert. Der praxisspezifische Effekt ist jeweils der Durchschnitt der Abweichungen (= mittlere Länge der farbigen Linien). Praxis «rot» weicht in allen AGG etwa gleich stark vom erwarteten Wert ab. Bei ihr ist die Streuung und somit der Unsicherheitsindikator gering. Anders ist die Situation bei Praxis «blau»: Ihre Abweichungen schwanken stark

und haben nicht alle das gleiche Vorzeichen. Der Unsicherheitsindikator ist bei Praxis «blau» wesentlich höher als bei Praxis «rot». Es ist aber möglich, dass die durchschnittliche Abweichung und somit der praxisspezifischen Effekt ($\hat{a}_i$) in beiden Praxen gleich ist.

Abbildung 2 Illustration des Unsicherheitsfaktors



Schwarze Punkte: Erwarteter Wert pro AGG bei gegebenen Morbiditätsvariablen (=Regressionsgerade)
Rote senkrechte Linien: Abweichung der Praxis «rot» vom erwarteten Wert
Blaue senkrechte Linien: Abweichung der Praxis «blau» vom erwarteten Wert

Die Abbildung illustriert den Unsicherheitsfaktor. Die schwarzen Punkte stellen die erwarteten Werte einer Praxis bei gegebenen Morbiditätsvariablen dar (Regressionsgerade), die senkrechten farbigen Linien sind die Abweichungen der beobachteten Werte vom erwarteten Wert (Residuen). Der praxisspezifische Effekt ist die durchschnittliche Abweichung (mittlere Länge der farbigen Linien). Der Unsicherheitsfaktor reflektiert die Streuung der Abweichungen um ihren Mittelwert. Praxis «rot» weicht in allen AGG etwa gleich ab. Die Streuung um den Mittelwert ist gering. Bei Praxis «blau» hingegen unterscheiden sich die Abweichungen in den einzelnen AGG stark. Der Unsicherheitsfaktor ist hoch.

Quelle: Eigene Darstellung, Polynomics.

Nicht berücksichtigt wird im Unsicherheitsindikator die Unsicherheit der gesamten Schätzung der Beta-Koeffizienten für die AGG und die Morbiditätsvariablen (in Abbildung 2 ist dies die schwarze Regressionslinie). Wir halten dies für ein valides Vorgehen, denn der Fokus liegt auf der Unsicherheit des *praxisspezifischen* Effekts.[2]

Erwähnenswert ist zudem, dass dieses Vorgehen von der Berechnung von cluster-robusten Standardfehlern abweicht, wie sie beispielsweise zur Beurteilung der statistischen Signifikanz von erklärenden Variablen auf der 1. Stufe im Modell (Gleichung (1)) verwendet werden könnten. Cluster-robuste Standardfehler berücksichtigen sowohl Heteroskedastie als auch die *Autokorrelation* innerhalb eines Clusters. Im vorliegenden Fall ist der Cluster die Praxis. Soll bei-

[2] Das Vorgehen ist vergleichbar dazu, wenn in einer empirischen Studie ausschliesslich die Standardfehler derjenigen Beta-Koeffizienten betrachtet werden, welche für die Studienresultate interpretiert werden sollen. Würde man die Fixed-Effects-Schätzung mit einem individuellen Dummy pro Arzt berechnen, und heteroskedastie-robuste Standardfehler für die Dummy-Variablen berechnen, erhielte man ähnliche Standardfehler wie mit dem hier vorgeschlagenen Unsicherheitsindikator. Das Vorgehen ist jedoch äusserst rechenintensiv, besonders bei Facharztgruppen mit vielen Ärzten

spielsweise der Standardfehler für den Indikator «Franchise» berechnet werden, gehen in diese Berechnung Daten von allen Arztpraxen ein. Deshalb kann berücksichtigt werden, dass unterschiedliche Praxen sich möglicherweise in der Streuung der Fehlerterme unterscheiden.

In die Berechnung des Standardfehlers für einen praxisspezifischen Effekt $\hat{a}_i$ fliessen jedoch nur die Werte *einer* Praxis ein (für alle anderen Praxen gilt ja ein anderer praxisspezifischer Effekt). Es steht hier also nur ein Cluster zur Verfügung und es existiert keine Komponente in der Streuung der Fehlerterme, welche sich zwischen den Clustern unterscheiden könnte. Die Berechnung von aussagekräftigen cluster-robusten Standardfehlern ist daher nicht möglich. Würde man in dieser Situation die übliche Formel für cluster-robuste Standardfehler auf den praxisspezifischen Effekt anwenden, würde dies zu viel zu kleinen Standardfehlern führen, weil sich innerhalb einer Praxis negative und positive Abweichungen vom Mittelwert gegenseitig aufheben (müssen). In der Literatur wird denn auch empfohlen, cluster-robuste Standardfehler erst ab einer Mindestanzahl von 50 Clustern zu berechnen (Cameron und Miller 2015).

4.4.2 Einbezug des Unsicherheitsindikators bei der Indexbildung

Der Unsicherheitsindikator $\sqrt{\frac{1}{J_i}\,\hat{\sigma}^2_{\hat{\omega}_{ij}}}$ kann dazu verwendet werden, einen Vertrauensbereich für den praxisspezifischen Effekt zu berechnen. Im vorliegenden Fall interessiert insbesondere die untere Grenze, weil der Fokus darauf liegt, nicht zu viele Praxen fälschlicherweise als auffällig zu identifizieren. Wir berechnen diese Untergrenze, indem wir vom Punktschätzer 1.96-mal den Unsicherheitsindikator abziehen. Dieses Vorgehen folgt der üblichen Berechnung eines 95-Konfidenzintervalls unter der Annahme, dass die Fehlerterme annähernd normalverteilt sind.

$$\hat{a}_i^{low} = \hat{a}_i - 1.96 * \sqrt{\frac{1}{J_i}\,\hat{\sigma}^2_{\hat{\omega}_{ij}}} \tag{9}$$

Aus der Untergrenze für den praxisspezifischen Effekt kann eine Untergrenze für den Indexwert einer spezifischen Praxis berechnet werden. Dazu wird in Gleichung (3) die Untergrenze des Effektes $\hat{a}_i^{low}$ anstatt der Punktschätzer $\hat{a}_i$ eingesetzt. Ansonsten wird an der Indexberechnung nichts verändert, auch nicht der Skalierungsfaktor. Es geht also bei der Untergrenze nicht darum, einen neuen Index zu berechnen, dessen Mittelwert selbst wieder 100 ist. Vielmehr soll pro Praxis zum bestehenden Index ein Vertrauensintervall nach unten berechnet werden.

4.5 Transformation der Zielvariablen

Die von Kaiser (2016) vorgeschlagene logarithmische Transformation der Zielvariable bringt klare Vorteile für die Modellberechnung. Untransformiert wäre die Zielvariable (Gesundheitsausgaben) stark rechtsschief verteilt. Auch bei einer rechtsschiefen Verteilung liefert die Methode der kleinsten Quadrate (Ordinary Least Squares, OLS) konsistente Schätzer, mit steigender Anzahl Beobachtungen streben die geschätzten Koeffizienten also immer näher an die wahren Werte. In kleineren Stichproben können jedoch die geschätzten Koeffizienten unpräzise sein. Zudem ist es möglich, dass Koeffizienten durch einzelne sehr hohe Werte (Ausreisser) nach oben «gezerrt» werden und somit den echten Einfluss überschätzen (Mihaylova et al. 2011). Die logarithmische Transformation führt zu annähernd normalverteilten Residuen (siehe empirische Auswertungen in Kapitel 11.3 im Anhang). Zudem kann ein Index als relative Grösse direkt berechnet werden.

Die logarithmische Transformation bringt indes auch Nachteile, besonders wenn das Ziel ist, am Ende eine Prädiktion in absoluten Franken zu haben. Auf diese Probleme werden wir in Kapitel 11.1 im Anhang eingehen. In vielen praktischen Anwendungen ziehen es die Autoren vor, ein untransformiertes Kostenmodell zu schätzen, besonders wenn das Ziel der Analyse die Prognose von Kosten ist (Buntin und Zaslavsky, 2004; Beck, 2013). Aus unserer Sicht kann rein aus theoretischen Gesichtspunkten nicht abschliessend gesagt werden, ob eine Logarithmierung der Zielvariablen angezeigt ist oder nicht. Wir werden in der empirischen Analyse daher zusätzlich zwei Modelle testen: ein untransformiertes Modell ohne Ausreisserkorrektur und ein Modell mit einer Ausreisserkorrektur (Winsorisierung). Ein empirischer Vergleich befindet sich in Abschnitt 11.3 im Anhang.

4.6 Einbezug weiterer Morbiditätsindikatoren

Ein Hauptziel der vorliegenden Analyse ist es, den Einbezug von weiteren Morbiditätsindikatoren in das statistische Screening zu prüfen. Dies steht auch im Einklang mit der Fachliteratur. In allen uns bekannten Artikeln zum «Physician Profiling» werden Indikatoren des Gesundheitszustands eingesetzt (siehe Abschnitt 3.1). Auch in der Schweiz wurde der Einbezug von weiteren Morbiditätsindikatoren schon von verschiedenen Gutachten gefordert (Schwenkglenks 2010; Wasem, Lux und Dahl 2010).

In einem ersten Schritt müssen Indikatoren ausgewählt werden, welche im Folgenden empirisch überprüft werden. Geeignete Indikatoren sollen die folgenden drei Kriterien erfüllen (Kaiser 2016, Abschnitt 3.1):

- **Erklärungsgehalt/Kostenrelevanz:** Die Morbiditätsindikatoren sollen einen starken Einfluss auf die Kosten haben.
- **Exogenität:** Die Morbiditätsindikatoren sind möglichst wenig durch das Arztverhalten beeinflussbar (keine Fehlanreize oder «Moral Hazard»).
- **Datenverfügbarkeit und -qualität:** Es können aus den verfügbaren Rohdaten valide Morbiditätsindikatoren zum Einbezug in das Modell gebildet werden.

In Tabelle 2 sind mögliche Morbiditätsvariablen aufgelistet und nach den Kriterien Erklärungsgehalt, Exogenität und Datenverfügbarkeit beurteilt. In der letzten Spalte ist angegeben, ob der Indikator in den empirischen Analysen untersucht wird oder nicht.

Der direkteste und in der Literatur am meisten verbreitete Morbiditätsindikator des Patientenstamms sind diagnostische Informationen (siehe z. B. Thomas et al., 2004a; sowie Adams et al., 2010b). In den Abrechnungsdaten der Schweizer Krankenversicherung stehen solche jedoch nicht zur Verfügung. Der Gesundheitszustand muss daher über Stellvertreterindikatoren («Proxies») angenähert werden.

Ein guter Proxy für den Gesundheitszustand ist die gewählte *Franchisestufe*. Eine hohe Franchisestufe lohnt sich vorwiegend für gesunde Personen, und mehrere Studien haben aufgezeigt, dass die durchschnittlichen Kosten von Personen mit hohen Franchisen deutlich geringer sind als von Personen mit der ordentlichen Franchise (Gardiol, Geoffard und Grandchamp 2005; Schmid und Beck 2015). Die Franchisestufe ist zudem wenig durch den Arzt beeinflussbar (exogene Variable).

Ähnlich wie die Franchisestufe werden auch gewisse *Versicherungsmodelle* (z. B. telemedizinische Angebote) vorwiegend von gesünderen Personen gewählt. Für einen Einbezug in das Modell ist dieser Indikator aus unserer Sicht jedoch trotzdem nicht geeignet, dies aus zwei haupt-

sächlichen Gründen: Erstens ist das Versicherungsmodell nicht unabhängig vom Arztverhalten. Falls Praxen in bestimmten Versicherungsmodellen risikobereinigt geringere Kosten aufweisen, wofür es in der Fachliteratur Evidenz gibt (Kauer 2016), soll sich dies auch in ihren Indexwerten widerspiegeln. Zweitens ist die Variable in den Daten schwer zu erfassen: Je nach Versicherer können sich die Modelle substantiell unterschieden, auch wenn sie ähnlich Namen haben.

Tabelle 2 Mögliche Morbiditätsindikatoren

	Erwarteter Erklärungsgehalt	Exogenität	Datenverfügbarkeit	Empirischer Test
Franchisen	Gut, in Studien belegt	Gut, geringer endogener Anteil	Gut in Arztdaten, sehr gut in Individualdaten	Ja
Versicherungsmodell	Mittel, abhängig vom konkreten Modell	Nicht unabhängig vom Arztverhalten	Vergleichbarkeit unklar	Nein
Spital-im-Vorjahr	Unklar, in Studien belegt für die Gesamtkosten der OKP	Mittel, beschreiben Inanspruchnahme	Sehr gut	Ja
Pharmazeutische Kostengruppen (PCG)	Sehr gut, in Studien belegt	Mittel, beschreiben Inanspruchnahme, problematisch bei medikamentösen Therapien, die substituierbar sind	Gut in Arztdaten, sehr gut in Individualdaten	Ja
Charakteristika des Praxisstandortes	Unklar	Gut, geringer endogener Anteil	Eingeschränkt durch Datenschutz[1)]	Ja
Kanton	Ohne Bereinigung der Preiseffekte: deutlich; mit Bereinigung der Preiseffekte: unklar	Sehr gut	Eingeschränkt durch Datenschutz[1)]	Ja
Informationen aus dem Tarmed	Bei spezifischen Arztgruppen hoch, z. B. zu Identifikation von Praxisauffälligkeiten wie Operationstätigkeit	Abhängig von der Ausgestaltung	Gut, zur Umsetzung detaillierte Spezifikation notwendig	Möglicherweise im dritten Teilprojekt
Diagnostische Kostengruppen	Sehr gut, in Studien belegt	Gut, Entscheidungsspielraum vorhanden	Nicht vorhanden	Nein

[1)] Diese Einschränkung ist relevant für Polynomics als externe Validierungsstelle. Für die Tarifpartner, welche die Wirtschaftlichkeitsprüfung umsetzten, ist sie weniger relevant.

In diesem Arbeitsschritt werden Morbiditätsindikatoren ausgewählt, welche in der späteren empirischen Analyse überprüft werden. Die Morbiditätsindikatoren werden nach den Kriterien «Erwarteter Erklärungsgehalt», «Exogenität» und «Datenverfügbarkeit» beurteilt. Zum empirischen Test in dieser Stufe vorgeschlagen sind die Indikatoren Franchisen, Spital-im-Vorjahr, pharmazeutische Kostengruppen sowie Standortvariablen.

Quelle: Eigene Darstellung.

Die beiden Indikatoren «*Spital-im-Vorjahr*» und «*pharmazeutische Kostengruppen*» wurden in der Risikoausgleichsliteratur schon seit vielen Jahren diskutiert (Beck 2013). Spital-im-Vorjahr ist seit 2012 auch im Risikoausgleich implementiert, pharmazeutische Kostengruppen werden es Jahr 2020. Bei beiden Indikatoren ist nachgewiesen, dass sie – zumindest in Bezug auf die Gesamtkosten zu Lasten der obligatorischen Krankenversicherung – substantielle Beiträge zur Kostenprognose leisten können.

Beide Indikatoren Spital-im-Vorjahr und pharmazeutische Kostengruppen werden aus der früheren Leistungsinanspruchnahme gebildet. Liegen keine diagnostischen Informationen vor, ist dies die nächstbeste Alternative, um den Leistungsbedarf des Patientenstamms abzubilden. Die Bildung von Morbiditätsvariablen aus der Inanspruchnahme ist indes nicht ohne Gefahren. Erstens ist die Inanspruchnahme nicht unabhängig vom Arztverhalten. So kann eine hohe Inanspruchnahme in der Vergangenheit sowohl durch einen besonders kranken Patientenstamm als auch durch andere Faktoren verursacht sein. Zweitens könnte es zu Fehlbeurteilungen führen, wenn Therapien, welche zur Bildung von «Morbiditätsfaktoren» verwendet werden, durch andere Therapien substituierbar sind. Falls beispielsweise durch eine intensive ärztliche Betreuung eine medikamentöse Therapie verhindert werden kann, besteht bei der Berücksichtigung von pharmazeutischen Kostengruppen die Gefahr, dass Praxen, welche eher auf die nichtmedikamentöse Therapie setzen, ungerechtfertigt «streng» beurteilt werden, denn ihr Patientenstamm wird als gesünder eingestuft, als er in Wahrheit ist. Leider kann mit statistischen Mitteln nicht beurteilt werden, wie gross die entsprechende Gefahr bezüglich von pharmazeutischen Kostengruppen ist. Dies muss bei der inhaltlichen Interpretation und Modellbeurteilung beachtet werden. Die inhaltliche Beurteilung könnte dazu führen, bestimmte pharmazeutische Kostengruppen nicht zu berücksichtigen, auch wenn sie einen statistisch signifikanten Einfluss auf die Kosten haben.

5 Datenbasis und Aufbereitungen

Im folgenden Abschnitt beschreiben wir die Bildung der Zielvariablen, sowie die Operationalisierung der ausgewählten Morbiditätsindikatoren mit den Daten des Sasis Tarif- oder Datenpools. Wir haben uns dabei auf die für das Modell wichtigsten Aspekte der Datenaufbereitung beschränkt. Zusätzliche Informationen zu Aggregationen und Verknüpfungen sind im Kapitel 11.2 im Anhang enthalten.

Die Daten des Sasis Tarif- oder Datenpools enthalten keinen Identifikator für einzelne Patienten.[3] Sie sind auf unterschiedlichen Ebenen voraggregiert: Die Leistungsdaten sind aggregiert nach Altersgruppe, Geschlecht, Franchise, Versicherungsmodell und Schadensart;[4] die Erkranktendaten sind aggregiert nach Praxis, Altersgruppen, Geschlecht und Spital-im-Vorjahr; und die Daten zur Bildung der pharmazeutischen Kostengruppen sind aggregiert nach Praxis, Alters- und Geschlechtsgruppe.

Um die Datensätze verknüpfbar zu machen, aggregieren wir alle Datensätze auf der Ebene der Praxis und Alters- und Geschlechtsgruppen (AGG). Die Information zu Franchisen und Spital-im-Vorjahr fällt indes nicht weg, sondern wird zur Bildung der Morbiditätsvariablen eingesetzt (siehe Abschnitt 5.2). Die Variable Versicherungsmodell werden wir im Modell nicht berücksichtigen (siehe Abschnitt 4.6). Den Informationsgehalt der Variablen Schadensart (Unterscheidung zwischen Krankheit, Unfall und Mutterschaft) schätzen wir als gering ein, so dass sie nicht zwingend berücksichtigt werden muss.

5.1 Bildung der Zielvariablen

5.1.1 Zuordnung der Leistungen zu den anonymisierten ZSR

Zur Bildung der Zielvariablen müssen die Kosten den Arztpraxen, welche mit anonymisierten ZSR-Nummern identifiziert sind, zugordnet werden. Dazu stehen die beiden Datenspalten «Verursacher» oder «Rechnungssteller» zur Verfügung. Dem Rechnungssteller werden Kosten für Leistungsarten zugeordnet, welche direkt in der Praxis erbracht werden. Sie werden auch als *direkte* Kosten bezeichnet. Sie sind in Tabelle 3 aufgelistet.

Die nach Tarmed abgerechneten Arztbehandlungen machen mit Abstand den grössten Teil der direkten Leistungsarten aus. Dabei ist zu beachten, dass sie in *Taxpunkten* und nicht in Franken gemessen werden. Dies trägt dazu bei, dass die erbrachte Leistungsmenge über die Kantone hinweg vergleichbar ist. Eine Beurteilung davon, ob die unterschiedlichen Taxpunktwerte gerechtfertigt sind, ist nicht Teil dieser Studie.[5]

Zu beachten ist zudem, dass bei den direkten Kosten jeder Praxis die Menge zugerechnet wird, welche von der Praxis selbst in Rechnung gestellt wurde. Wenn ein Arzt einen Patienten an einen zweiten Arzt weiterverweist, werden beim ersten Arzt nur die selbst erbrachten Leistungen gezählt, die Leistungen des zweiten Arztes zählen beim zweiten Arzt. Diese Zuordnung hat einerseits datentechnische Gründe: Da Patienten in der Schweiz keine Überweisungen brauchen,

---

[3] Für die Auswertungen auf der Basis von Individualdaten siehe Kapitel 9.

[4] Technisch existiert die «Leistungsart» als zusätzliche Gruppierungsebene. Diese gilt für uns aber nicht als Gruppierung, sondern wird gebraucht um die Zielvariable zu bilden (siehe Abschnitt 5.1.1).

[5] Bei einer operativen Umsetzung des Modells wäre jedoch zu diskutieren, ob eine Gewichtung mit einem einheitlichen Taxpunktwert (z. B. 0.88 als gewichteter Durchschnitt aller in der Schweiz eingesetzter Taxpunktwerte) nicht angebrachter wäre als eine Gewichtung mit dem Wert eins. Die Gewichtung mit dem Wert eins bedeutet, dass Arztleistungen gegenüber anderen Leistungen (z. B. Medikamenten) verteuert werden.

wenn sie unterschiedliche Ärzte aufsuchen, sind Weiterweisungen zwischen den Ärzten in den Abrechnungsdaten nicht durchgehend erfasst. Andererseits ist es aber auch inhaltlich unklar, inwiefern ein veranlassender Arzt Kontrolle über die Kosten bei anderen Ärzten hat.

Tabelle 3 Direkte Kosten werden dem anonymisierten Rechnungssteller zugeordnet

	Anzahl Beobachtungen	Anzahl Rechnungssteller	Summe Wert (CHF)[a)]
Arztbehandlung Tarmed	12'486'072	21'912	7'233'188'488
Arzt ambulant/Medikamente	4'878'863	16'486	1'821'441'960
Arzt ambulant/Analysen	4'170'960	13'057	503'105'202
Arzt ambulant/übrige Tarife	1'457'463	14'684	133'859'594
Arzt ambulant/MiGeL	796'714	18'921	173'098'286
Arzt ambulant/Physiotherapie	6'199	1'571	2'055'681
Gesamt	**23'796'271**		**9'866'749'211**

Die Zahlen entsprechen dem Datenjahr 2015. Es sind nur Rechnungssteller mit mind. einem Patienten im Erkranktenrecord dargestellt.

[a)] Bei den Tarmedleistungen sind es Taxpunkte.

In der Tabelle sind die Leistungsarten aufgelistet, bei welchen die Kosten der anonymisierten ZSR-Nummer des Rechnungsstellers zugerechnet werden («Direkte Kosten»). Darunter fallen alle Tarmedleistungen. Bei einer Überweisung von einem Arzt zum anderen, wird jedem Arzt die Leistungsmenge zugerechnet, welche er selbst erbracht hat.

Quelle: Daten aus dem Sasis Datenpool, Leistungsrecord 2015; eigene Berechnungen.

In Tabelle 4 sind die Leistungsarten aufgelistet, deren Kosten dem veranlassenden Arzt zugerechnet werden. Dies sind Leistungsarten, welche Patienten in der Schweiz nicht ohne ärztliche Verschreibung zu Lasten der obligatorischen Krankenversicherung abrechnen können. Den grössten Teil dieser Leistungen machen die Medikamente aus, gefolgt von Laborleistungen und Physiotherapie.

Tabelle 4 Indirekte Kosten werden dem anonymisierten Veranlasser zugeordnet

	Anzahl Beobachtungen	Anzahl Veranlasser	Summe Wert (CHF)
Apotheken/Medikamente A&B, übrige pflicht-med. Taxen, Analysen	6'320'434	21'230	2'398'443'794
Laboratorien/Analysen	3'216'904	16'795	693'527'971
Physiotherapeuten/Physiotherapie	1'023'183	15'434	650'295'932
Abgabestelle/MiGeL	336'314	13'640	228'261'054
Apotheken/MiGeL	628'644	16'409	121'615'503
Gesamt	**11'525'479**		**4'092'144'254**

Die Zahlen entsprechen dem Datenjahr 2015. Es sind nur Veranlasser mit mind. einem Patienten im Erkranktenrecord dargestellt.

Die hier dargestellten Leistungsarten («indirekte Kosten») werden der anonymisierten ZSR-Nummer des veranlassenden Leistungserbringers zugerechnet. Den mit Abstand grössten Anteil machen die Medikamentenkosten aus.

Quelle: Daten aus dem Sasis Datenpool, Leistungsrecord 2015, Eigene Berechnungen.

5.1.2 Aggregation, Logarithmierung und Winsorisierung

Grundsätzlich ist die Zielvariable der Schätzung der Durchschnitt der Kosten pro Praxis und AGG. Pro Praxis stehen im Mittelwert 35 Beobachtungen zur Verfügung, weniger sind es bei den Facharztgruppen Gynäkologie sowie Kinder- und Jugendmedizin (durchschnittlich 19 Beobachtungen).

Wie in Abschnitt 4.5 erwähnt, testen wir drei unterschiedliche Spezifikationen dieser Zielvariablen: Die logarithmierte Transformation (Kaiser 2016), die untransformierten Kosten und die «Winsorisierung». Bei der Logarithmierung können keine Kosten berücksichtigt werden, welche null oder negativ sind. Da solche Werte auf Stornierungen zurückzuführen sind, werden sie in allen Schätzungen ausgeschlossen.

Die Winsorisierung wird pro Facharztgruppe auf dem 95-Prozent-Perzentil vorgenommen. Das 95-Prozent-Perzentil ist der Wert, bei welchem 95 Prozent der beobachteten Kosten darunterliegen und 5 Prozent der Kosten darüber. Beobachtete Werte, die über dem 95-Prozent-Perzentil liegen, werden auf den Wert des 95-Prozent-Perzentils «heruntergestutzt». Sie werden also nicht ausgeschlossen und haben weiterhin einen hohen Wert (wesentlich höher als der Rest der Verteilung). Sie gehen aber nicht mit ihrer vollen Grösse in die Schätzung ein.

5.2 Bildung der Morbiditätsindikatoren

5.2.1 Bildung der Indikatoren «Franchise» und «Spital-im-Vorjahr»

In den Sasisdaten ist nicht direkt identifizierbar, wie viele Erkrankte welche Franchisen haben. Im Leistungsrecord steht jedoch pro Praxis, AGG und Franchisestufe die Anzahl Konsultationen zur Verfügung. Pro Praxis und AGG haben wir daraus den Anteil an Konsultationen berechnet, welche für Versicherte mit hohen Franchisen stattfanden. Als hohe Franchisen gelten alle Franchisestufen von mindestens CHF 1‘000 bei Erwachsenen und CHF 200 bei Kindern. Die Variable hat das Format einer kontinuierlichen Variablen und nimmt Werte zwischen null und eins an.

Den Indikator Spital-im-Vorjahr, welcher seit 2012 im Risikoausgleich verwendet wird, spezifizieren wir in einer ähnlichen Art wie die Wahlfranchisen. Pro Praxis und AGG berechnen wir, welcher Anteil der Patienten im Vorjahr einen Spitalaufenthalt aufwies. Auch dies ergibt eine kontinuierliche Variable mit Werten zwischen null und eins.

5.2.2 Bildung der pharmazeutischen Kostengruppen

Zur Bildung der pharmazeutischen Kostengruppen haben wir eine PCG-Klassifikation verwendet, welche für den Risikoausgleich zwischen den Krankenversicherern entwickelt wurde. Diese stellt eine Adaption der pharmazeutischen Kostengruppen dar, welche im Jahr 2014 im niederländischen Risikoausgleich eingesetzt wurden (Trottmann et al. 2015). Sie enthält 24 pharmazeutische Kostengruppen.

Die Klassifikation wurde nicht speziell für die Wirtschaftlichkeitsprüfungen entwickelt; dies vor allem aus Ressourcengründen: Eine eigene Weiterentwicklung wäre im sehr engen zeitlichen und finanziellen Rahmen dieses Projektes nicht möglich gewesen. Es ist jedoch denkbar, dass durch spezifische PCG pro Facharztgruppe deutliche Verbesserungen am Erklärungsgehalt der Modelle erreicht werden könnten.

Für die operative Umsetzung der PCG-Klassifikation mit den bestehenden Daten haben wir die folgenden Schritte unternommen.

Berechnung der Wirkstoffmenge in standardisierten Tagesdosen (DDD-Menge)

Für die Eingruppierung in die PCG muss die Wirkstoffmenge von unterschiedlichen Medikamenten vergleichbar gemacht werden. In der PCG-Klassifikation des Risikoausgleichs werden dazu sogenannte «standardisierte Tagesdosen» (Defined Daily Dosis, DDD) eingesetzt. Diese von der WHO eingeführte Masszahl[6] gibt pro Wirkstoff und Darreichungsform eine «durchschnittliche Dosis pro Tag in der Hauptindikation für Erwachsene» an. Dies ist keinesfalls eine Behandlungsempfehlung und die Dosierung für individuelle Patienten kann deutlich davon abweichen. Für statistische Zwecke hat sich die Masszahl aber bewährt.

Die PCG-Liste enthält pro PCG eine Gruppe an Wirkstoffen (ATC-Code), die dieser PCG zugeordnet sind, und zu diesen Wirkstoffen ihre Darreichungsformen und die DDD pro Darreichungsform. In Abbildung 3 zeigen wir schematisch die Zuordnung eines Pharmacodes zu einer PCG und die Berechnung der DDD-Menge. Der Pharmacode enthält 4.5 mg des Wirkstoffes Indacaterol (ATC-Code: R03AC18) als Pulverinhalation (Inhal.powder). Über den ATC-Code und die Darreichungsform wird der Pharmacode mit der PCG-Liste verknüpft. Der ATC-Code ist in der PCG «COPD/Schweres Asthma» (PCG 2), und die DDD für die Darreichungsform «Pulverinhalation» beträgt 0.15 mg. Eine Packung mit 4.5 mg enthält demnach 30 DDD.

Abbildung 3 PCG-Zuordnung und Berechnung der DDD-Menge

(1) Pharmacode bzw. GTIN-Code

Pharmacode	Produktname	AtcCode	AdmCode	Wirkstoffmenge pro Packung
4603089	Onbrez Breezhaler, Inh Kaps 0.150 mg, 30 Stk	R03AC18	Inhal.powder	4.5 mg

Verknüpfung

(2) PCG-Zuordnung

PCG-Bezeichnung	PCG-Nr	AtcCode	AdmCode	DDD
COPD/Schweres Asthma	2	R03AC18	Inhal.powder	0.15 mg

Berechnung: 4.5/0.15

(3) Eingruppierung des Pharmacode bzw. GTIN-Code

Pharmacode	PCG-Nr	DDD-Menge pro Packung
4603089	2	30

In der Abbildung ist beispielhaft die Eingruppierung einer Medikamentenpackung (Pharmacode) in eine PCG gezeigt. Für die Eingruppierung wird der Pharmacode mit der PCG-Liste verknüpft. Die Verknüpfungsvariablen sind der ATC-Code und die Darreichungsform Die PCG-Liste ihrerseits ordnet den ATC-Code einem Indikationsgebiet (PCG) zu und bestimmt Wirkstoffmenge, welche für diesen ATC-Code und für diese Darreichungsform als standardisierte Tagesdosis (DDD) gelten. Daraus kann die DDD-Menge pro Pharmacode berechnet werden.

Quelle: Eigene Darstellung, Polynomics.

[6] Siehe: https://www.whocc.no/ddd/definition_and_general_considera/, abgerufen am 25.9.2017.

In den Daten aus dem Sasis Tarifpool sind die einzelnen abgerechneten Pharmacodes pro Praxis und AGG verfügbar. Diese haben wir mit der PCG-Liste verknüpft und daraus die gesamte DDD-Menge berechnet, welche pro PCG, Praxis und AGG abgerechnet wurde.

Ausschluss spezifischer PCG pro Facharztgruppe

Eine bestimmte PCG wird für eine Facharztgruppe nur dann berücksichtigt, wenn mehr als 30 Ärzte innerhalb der Facharztgruppe eine Mindestmenge oder mehr Medikamente aus der entsprechenden PCG veranlasst haben. Die Mindestmenge, damit eine Praxis zählt, wurde dabei auf geringe 1.8 DDD festgelegt. Dies entspricht einer Halbjahresdosis (180 definierte DDD) pro 100 Patienten.

Der Grund für die Einschränkung auf 30 Ärzte pro Facharztgruppe liegt in der Stabilität der geschätzten Koeffizienten. Falls es pro Facharztgruppe nur wenige Beobachtungen mit Medikamenten aus der spezifischen PCG gibt, wäre es problematisch, die Gruppe zu berücksichtigen. Die geschätzten Koeffizienten könnten zu stark von einzelnen Beobachtungen geprägt sein und keinen generellen Zusammenhang beschreiben.

Hierarchisierung

Bei der Anwendung der PCG im niederländischen Risikoausgleich erfolgt eine «Hierarchisierung» der PCG. Verwandte Indikationsgebiete werden zu einer Hierarchie zusammengefasst (Beispiel: Asthma und COPD). Erfüllt ein Patient die Aufnahmekriterien für mehrere PCG in der gleichen Hierarchie, wird er nur der schwersten PCG zugeordnet (Beispiel: Ein Patient, der sowohl Medikamente aus dem Indikationsgebiet Asthma als auch COPD bezogen hat, wird nur der Gruppe COPD zugeteilt). Diese Hierarchisierung wurde in den Niederlanden vorgenommen, um Manipulationsmöglichkeiten einzuschränken, denn es wird über den Risikoausgleich sehr viel Geld umverteilt. Statistisch gesehen bringt die Hierarchisierung eher eine Verschlechterung: Prädiktionsmodelle ohne diese Einschränkung haben einen besseren Erklärungsgehalt (Trottmann et al. 2015).

Eine Umsetzung der Hierarchisierung mit aggregierten Daten ist nicht möglich. Es wurde daher jede PCG für sich alleine berücksichtigt. Eine Beobachtung kann beispielsweise sowohl bei der PCG «Asthma» als auch bei der PCG «COPD» einen positiven Wert haben. Eine inhaltliche Änderung ergibt sich durch die Weglassung der Hierarchisierung bei PCG 23 (Diabetes Typ II mit Bluthochdruck) und PCG 8 (Diabetes Typ II ohne Bluthochdruck). Im Risikoausgleich kann ein Patient nur dann in PCG 23 eingeteilt werden, wenn er gleichzeitigt die Kriterien für PCG 8 erfüllt. Da nicht beurteilt werden kann, welche Patienten Medikamente aus beiden Indikationsgebieten bezogen haben, werden die beiden PCG 8 (Diabetes Typ II) und 23 (Bluthochdruck) getrennt betrachtet. Dies führt zu einer deutlichen Erhöhung der Anzahl Patienten in PCG 23.

Berücksichtigung der Wirkstoffmenge

Wie bereits erwähnt haben wir in den aggregierten Daten die Wirkstoffmenge (DDD-Menge) pro PCG, Praxis und AGG berechnet. Diese DDD-Menge soll in der Regression als erklärende Variable verwendet werden. Für die konkrete Umsetzung gibt es mehrere Möglichkeiten. Erstens kann die DDD-Menge direkt als kontinuierliche Variable in das Modell aufgenommen werden. Damit wird implizit ein linearer Zusammenhang zwischen der verschriebenen DDD-Menge und den Kosten pro AGG und Praxis angenommen. Alternativ können Gruppen gebildet werden, die sich an der empirischen Verteilung der Wirkstoffmengen orientieren (umgangssprachlich ausgedrückt: es werden Gruppen gebildet für «geringe Wirkstoffmenge», «mittlere

Wirkstoffmenge» etc.). Diese zweite Alternative hat den Vorteil, dass auf die Annahme eines linearen Zusammenhangs verzichtet werden kann. Nachteilig ist jedoch, dass Schwelleneffekte auftreten können: Ist ein Wert nahe an einer Gruppengrenze, führt eine kleine Änderung der Wirkstoffmenge zu einer anderen Eingruppierung und somit zu einer anderen Kostenprognose.

In der empirischen Umsetzung haben wir uns für die zweite Möglichkeit entschieden. Bei der Gruppenbildung war es uns wichtig, dass in jeder Gruppe ausreichend Beobachtungen zur Verfügung stehen. Die Anzahl Gruppen wurde daher der Anzahl verfügbaren Beobachtungen pro PCG und Facharztgruppe angepasst. Die Schwellen zur Gruppeneinteilung wurden folgendermassen gewählt:

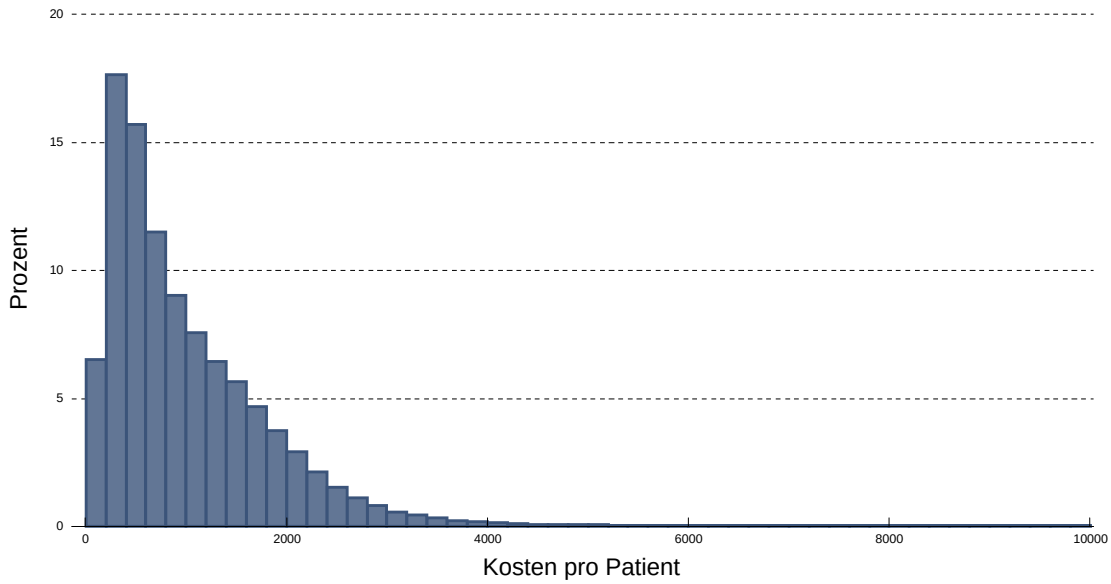
- Eigene Kategorie für DDD-Menge = 0.
- Falls es über 10’000 Beobachtungen mit einem Wert grösser null gibt, werden zehn Kategorien nach den Dezilen der Verteilung gebildet.
- Falls es 1’000 bis 10’000 Beobachtungen mit einem Wert grösser null gibt, werden vier Kategorien nach den Quartilen gebildet.
- Falls es 0 bis 1’000 Beobachtungen mit einem Wert grösser null gibt, werden zwei Kategorien gebildet. Der Split erfolgt beim Median.

6 Empirische Analysen, erste Stufe

6.1 Verteilung der Zielvariablen vor und nach der Transformation

Gesundheitsausgaben weisen typischerweise eine sehr rechtsschiefe Verteilung auf. Das heisst, es gibt sehr viele Beobachtungen mit sehr geringen Kosten und eine kleine Gruppe Beobachtungen mit sehr hohen Kosten. Abbildung 4 illustriert dies anhand der Verteilung der Kosten pro Patient für die Facharztgruppe «Allgemeine Innere Medizin». Der erste Balken, welcher Beobachtungen mit Kosten bis 200 Franken beinhaltet, umfasst rund 6 Prozent der Beobachtungen. Der Median, welcher die Beobachtungen in die 50 Prozent mit den höchsten und die 50 Prozent mit den geringsten Kosten teilt, liegt bei CHF 867 und somit deutlich unter dem Mittelwert von CHF 1'026. Die Schiefe liegt bei einem Wert von 6.45, was einer sehr deutlich rechtsschiefen Verteilung entspricht.

Abbildung 4 Untransformierte Kosten pro Patient, Allgemeine innere Medizin



N = 205‘430, 258 Beobachtungen mit Kosten > CHF 10‘000 nicht dargestellt.

Untransformiert sind die Gesundheitsausgaben stark rechtsschief verteilt, d. h. die meisten Beobachtungen haben geringe Kosten (50% der Beobachtungen haben Kosten bis 867 Franken pro Patient), während eine kleine Minderheit deutlich höhere Ausgaben aufweist.

Quelle: Daten der Sasis AG; Datenjahr 2015; eigene Berechnungen.

Das in Abschnitt 4.1 beschriebene Fixed-Effects-Modell wird mit der üblichen Regressionsmethode der kleinsten Quadrate (OLS) geschätzt. Diese Methode liefert unabhängig von der Verteilung der Zielvariablen konsistente Schätzer, mit steigender Anzahl Beobachtungen streben die geschätzten Koeffizienten also immer näher an die wahren Werte. In kleineren Stichproben jedoch können die geschätzten Koeffizienten unpräzise sein. Bei einer rechtschiefen Verteilung der Zielvariablen ist es dann möglich, dass Koeffizienten durch einzelne sehr hohe Werte (Ausreisser) nach oben gezerrt werden und somit den echten Einfluss überschätzen.

Eine erste Reduktion der Rechtsschiefe wird dadurch erreicht, dass die sehr hohen Kosten «gestutzt» werden (Winsorisierung). Konkret werden Werte, die über dem 95-Prozent-Perzentil der Verteilung liegen, auf den Wert des 95-Prozent-Perzentils heruntergesetzt. Sie werden also nicht ausgeschlossen und bleiben deutlich über dem Grossteil der Verteilung, sie sind aber nicht mehr ganz so extrem wie vor der Winsorisierung. Wie in Tabelle 5 gezeigt, vermindert sich bei den Allgemeinärzten durch die Winsorisierung der Mittelwert um rund CHF 20 oder rund 2 Prozent. Die Stutzung der Extremwerte reduziert also die Gesamtsumme der berücksichtigten Kosten um rund 2 Prozent. Die Schiefe der Verteilung geht deutlich zurück, auch die Standardabweichung sinkt.[7]

Tabelle 5 Verteilung der Zielvariablen, Allgemeine Innere Medizin

	N	Mittel-wert	Std. Abw.	p25	p50	p75	p95	Max.	Schiefe
Untransformiert	205'425	1'026	738	503	867	1'392	2'261	99'783	6.45
Winsorisiert	205'425	1'004	617	503	867	1'392	2'261	2'591	0.80
Logarithmiert	205'425	6.71	0.70	6.22	6.76	7.24	7.72	11.51	-0.41

Beobachtungen gewichtet mit der Anzahl Erkrankten.

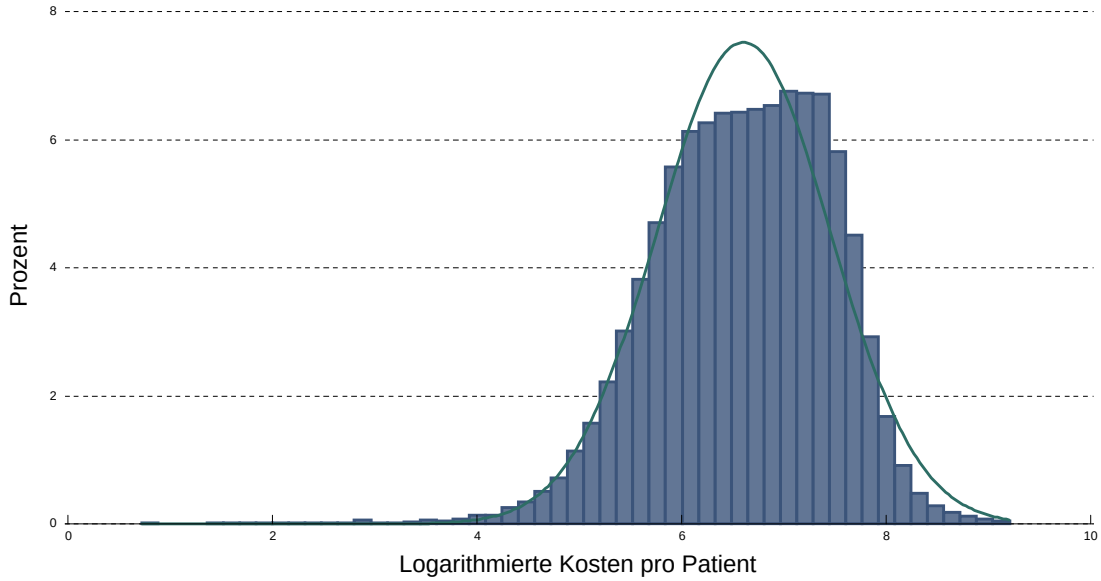
Untransformiert sind die Kosten sehr rechtsschief verteilt: Der Median liegt deutlich über dem Mittelwert, die Schiefe hat einen hohen Wert von über sechs und die teuersten 5% (Werte, die zwischen dem 95%-Perzentil und dem Maximum liegen) sind auf einer viel grösseren Spannweite verteilt als alle anderen Werte. Nach der logarithmischen Transformation ist die Verteilung deutlich symmetrischer. Die Schiefe wird sogar negativ, die Verteilung ist leicht linksschief.

Quelle: Daten der Sasis AG; Datenjahr 2015; eigene Berechnungen.

Der Effekt der logarithmischen Transformation ist in Abbildung 5 dargestellt, die grüne Linie stellt die Dichtefunktion der Normalverteilung dar. Deutlich erkennt man, dass die Verteilung nun annähernd symmetrisch ist. Die eher geringen Werte werden durch das Logarithmieren auseinandergezogen, die hohen Werte werden gestutzt. Auch die Kennzahlen der Verteilung weisen nun auf eine annähernd normalverteilte Variable hin. Der Median liegt nahe beim Mittelwert und die Schiefe ist sogar leicht negativ (leicht linksschiefe Verteilung).

[7] Es fällt auf, dass das Maximum im winsorisierten Modell nicht dem Wert des 95-Prozent-Perzentils der untransformierten Verteilung entspricht. Der Grund dafür ist, dass in der Darstellung die Werte mit der Summe Erkrankter gewichtet sind (wie sie auch in die Regression eingehen). Für die Winsorisierung wurde das 95-Prozent-Perzentil aber ungewichtet berechnet. Da Ausreisser häufig bei Werten mit wenigen Erkrankten vorkommen, liegt das ungewichtet berechnete 95-Prozent-Perzentil über dem gewichteten.

Abbildung 5 Allgemeine Innere Medizin: Logarithmierte Kosten pro Patient



N = 205‘172, 258 Beobachtungen mit Kosten > CHF 10‘000 nicht dargestellt.
Nach der logarithmischen Transformierung der Ausgangsdaten ist die Kostenverteilung annähernd symmetrisch. Die durchgezogene Linie zeigt zum Vergleich den Verlauf einer Normalverteilung.

Quelle: Daten der Sasis AG; Datenjahr 2015; eigene Berechnungen.

In Tabelle 6 ist die Verteilung für ausgewählte Facharztgruppen dargestellt. In den Facharztgruppe der Kinder- und Jugendmedizin ist die Schiefe extrem hoch. Dieser Extremwert kommt durch einige sehr grosse Ausreisser zustande. Es handelt sich um Beobachtungen mit sehr hohen Medikamentenkosten, welche zwar extrem sind, aber vorkommen können (keine zwingenden Datenfehler). Sie haben aber eine geringe Patientenzahl und werden in der Regression dementsprechend gering gewichtet.

Bei den Ärzten der Ophthalmologie werden durch die Winsorisierung die Gesamtkosten am stärksten reduziert, der Mittelwert im winsorisierten Modell ist um 16 Prozent geringer als der Mittelwert in den Originaldaten. Bei dieser Facharztgruppe liegen im obersten 5 Prozent der Verteilung also nicht nur «echte Ausreisser» (mit normalerweise wenigen Erkrankten), sondern ein beachtlicher Teil der Kosten. Möglicherweise führt eine Besonderheit im Leistungsspektrum bei einem Teil dieser Facharztgruppe zu deutlich anderen Kosten.

Tabelle 6 Verteilung der Zielvariablen für ausgewählte Facharztgruppen

	N	Mittel-wert	Std. Abw.	p25	p50	P75	p95	Max	Schiefe
Chirurgie									
Untransformiert	15'543	588	424	323	481	724	1'303	31'674	5.07
Winsorisiert	15'543	561	314	323	481	724	1'303	1'348	0.97
Logarithmiert	15'543	6.19	0.60	5.78	6.18	6.58	7.17	10.36	0.13
Gynäkologie									
Untransformiert	23'401	531	284	353	466	671	1'035	33'678	5.22
Winsorisiert	23'401	520	240	353	466	671	1'035	1'044	0.52
Logarithmiert	23'401	6.13	0.59	5.87	6.14	6.51	6.94	10.42	-1.28
Kardiologie									
Untransformiert	14'295	870	368	671	825	992	1'431	23'727	6.59
Winsorisiert	14'295	852	274	671	825	992	1'431	1'593	0.65
Logarithmiert	14'295	6.70	0.36	6.51	6.71	6.90	7.27	10.07	-0.30
Kinder- und Jugendmedizin									
Untransformiert	17'115	445	742	304	407	547	751	248'449	226.12
Winsorisiert	17'115	437	179	304	407	547	751	1'042	0.78
Logarithmiert	17'115	6.00	0.45	5.72	6.01	6.30	6.62	12.42	-0.47
Ophthalmologie									
Untransformiert	34'238	534	641	265	366	521	1'515	25'118	4.93
Winsorisiert	34'238	449	277	265	366	521	1'202	1'202	1.53
Logarithmiert	34'238	5.99	0.65	5.58	5.90	6.26	7.32	10.13	1.23
Psychiatrie und Psychotherapie									
Untransformiert	59'562	2'368	1'322	1'535	2'145	2'898	4'645	30'885	2.40
Winsorisiert	59'562	2'328	1'145	1'535	2'145	2'898	4'645	5'727	0.91
Logarithmiert	59'562	7.63	0.57	7.34	7.67	7.97	8.44	10.34	-0.89

Beobachtungen gewichtet mit der Anzahl Erkrankten.

Die Schiefe der Verteilung ist bei der Facharztgruppe Kinder- und Jugendmedizin am höchsten. In der Ophthalmologie hat die Winsorisierung einen starken Effekt auf die berücksichtigten Gesamtkosten, der Mittelwert sinkt um 15%. Ein beachtlicher Teil der Kosten befindet sich also «im Schwanz der Verteilung».

Quelle: Daten der Sasis AG; Datenjahr 2015; eigene Berechnungen.

6.2 Morbiditätsindikatoren

6.2.1 Alters- und Geschlechtsgruppen (AGG)

Bereits im heute verwendeten ANOVA-Verfahren werden Alters- und Geschlechtsgruppen zur Berechnung des Indexes eingesetzt. Auch wenn es sich nicht um direkte Indikatoren des Gesundheitszustands handelt, sind sie doch klar mit den Gesundheitsausgaben korreliert. Zur Bildung des Indikators werden jeweils fünf Jahre zusammengefasst, ab dem 96. Altersjahr gibt es nur noch eine Gruppe.

In Tabelle 7 sind Kennzahlen eines Regressionsmodells angegeben, welches nur AGG und einen praxisspezifischen Effekt als erklärende Variablen enthält. Mit einem korrigierten Bestimmtheitsmass (Adjusted R^2) von rund 70 bis 90 Prozent ist der Erklärungsgehalt gut. Am geringsten ist der Erklärungsgehalt der AGG auf die Kosten bei den Facharztgruppen der Kinder und Jugendmedizin sowie der Psychiatrie und Psychotherapie. Bei der Kinder- und Jugendmedizin ist dabei das R^2 mit anderen Modellen vergleichbar, der durchschnittliche absolute Prognosefehler (Mean Absolute Prediction Error, MAPE) ist jedoch deutlich grösser. Dies ist ein Hinweis dafür, dass es hohe Ausreisser gibt, diese jedoch durch die Modellvariablen recht gut erklärt werden. Das R^2 ist stark dadurch bestimmt, wie gut die Ausreisser erklärt werden, es ist daher recht hoch. Der MAPE, der nicht so stark von den Ausreissern bestimmt wird, hat einen vergleichsweise schlechteren Wert.

Tabelle 7 Erklärungsgehalt eines Modells mit nur AGG nach Facharztgruppen

	Allgemeine Innere	Chirurgie	Gynäkologie	Kardiologie	Kinder/ Jugend	Ophthal-mologie	Psychiatrie/ Psycho-therapie
N	205'430	15'543	23'401	14'295	17'115	34'238	59'562
N Ärzte	5‘455	481	1‘221	425	1‘020	871	2‘487
Adj. R^2	0.817	0.729	0.892	0.700	0.745	0.862	0.427
MAPE	0.330	0.347	0.292	0.246	0.508	0.259	0.478

Zielvariable: Log-transformiert, Gewichtung der Regression mit der Anzahl Erkrankten pro AGG
N: Stichprobengrösse; adj. R^2: korrigiertes Bestimmtheitsmass; MAPE: durchschnittlicher absoluter Prognosefehler.

Der Zusammenhang von Alter- und Geschlechtsgruppen mit den Gesundheitsausgaben ist klar gegeben. Ein Modell, welches nur AGG und einen spezifischen Praxiseffekt enthält, erreicht R^2-Werte von rund 70 bis 90% (mit aggregierten Daten). Der geringste Erklärungsgehalt wird bei der Facharztgruppe Psychiatrie und Psychotherapie sowie bei der Kinder- und Jugendmedizin erreicht.

Quelle: Daten der Sasis AG; Datenjahr 2015; eigene Berechnungen.

6.2.2 Indikator Franchisen

Eine hohe Franchise lohnt sich vorwiegend für Personen, die gesund sind und einen geringen Leistungsbedarf aufweisen. Mehrere Studien haben gezeigt, dass die durchschnittlichen Kosten von Personen mit hohen Franchisen deutlich geringer sind als von Personen mit der ordentlichen Franchise (Schmid und Beck, 2015; Gardiol et al., 2005).

In Tabelle 8 sind die Koeffizienten des Indikators für ausgesuchte Facharztgruppen dargestellt. Das ökonometrische Schätzmodell ist dabei das komplette Modell mit AGG, Franchisen, Spital-im-Vorjahr und PCG aus erklärenden Variablen. Die Zielvariable ist logarithmiert.

Bei den meisten Facharztgruppen zeigt sich der erwartete negative Zusammenhang zwischen dem Anteil an hohen Franchisen und den Kosten, in der Facharztgruppe Kardiologie ist er jedoch statistisch nicht signifikant. Ein kontraintuitiver, positiver Zusammenhang ist in der Facharztgruppe Gynäkologie zu beobachten. Die Gynäkologie ist insofern ein Sonderfall, als dass die Leistungen im Zusammenhang mit Mutterschaft von der Franchise ausgenommen sind. Auch für Patientinnen, welche einen hohen Bedarf an diesen Leistungen aufweisen, kann sich eine hohe Franchise durchaus lohnen.

Tabelle 8 Einfluss der Franchisen, Gesamtmodell nach Facharztgruppen

	Allgemeine Innere	Chirurgie	Gynäkologie	Kardiologie	Kinder/ Jugend	Ophthal-mologie	Psychiatrie/ Psycho-therapie
N	205'430	15'543	23'401	14'295	17'115	34'238	59'562
N Ärzte	5‘455	481	1‘221	425	1‘020	871	2‘487
Koeffizient Logmodell	-0.298** (0.00)	-0.069** (-0.00)	0.090** (0.01)	-0.018 (0.31)	-0.445*** (0.00)	-0.128*** (0.00)	-0.068*** (0.00)

Koeffizienten aus dem Gesamtmodell mit AGG, Franchisen, Spital-im-Vorjahr und PCG, Zielvariable logarithmiert. Standardfehler in Klammern. Signifikanzniveaus: *** p<0.01, ** p<0.05, * p<0.1.

In der Tabelle sind für ausgesuchte Facharztgruppen die Koeffizienten des Indikators Franchise dargestellt. Der Indikator ist in den meisten Facharztgruppen signifikant negativ. In der Facharztgruppe Gynäkologie ist ein signifikant positiver Effekt zu beobachten.

Quelle: Daten der Sasis AG; Datenjahr 2015; eigene Berechnungen.

Die Koeffizienten für die weiteren Facharztgruppen sind aus Platzgründen nicht dargestellt, sie sind im Excel Anhang «Polynomics_Wirtschaftlichkeitspruefungen_Regressionskoeffizienten_Stufe_1» verfügbar. Insgesamt ergaben sich signifikant negative Koeffizienten in 20 Facharztgruppen. In 10 Facharztgruppen waren die Effekte statistisch nicht signifikant, davon waren zwei insignifikant positiv. Die bereits erwähnte Facharztgruppe der Gynäkologen ist die einzige, in welcher ein signifikant positiver Zusammenhang beobachtet wurde.

6.2.3 Indikator Spital-im-Vorjahr

In Tabelle 9 sind die Koeffizienten des Indikators Spital-im-Vorjahr für ausgesuchte Facharztgruppen dargestellt. Der Indikator ist positiv signifikant für die Facharztgruppen Allgemein Innere Medizin, Chirurgie, Kinder und Jugendmedizin und Psychiatrie und Psychotherapie. Die grössten, und somit ökonomisch relevantesten Effekte finden sich bei Allgemein Innere Medizin, Kinder- und Jugendmedizin und Psychiatrie und Psychotherapie. Bei den anderen Facharztgruppen ist der Einfluss nicht signifikant, mit Ausnahme der Ophthalmologie jedoch positiv.

Tabelle 9 Einfluss der Variablen Spital-im-Vorjahr, Gesamtmodell nach Facharztgruppen

	Allgemeine Innere	Chirurgie	Gynäkologie	Kardiologie	Kinder/ Jugend	Ophthal-mologie	Psychiatrie/ Psycho-therapie
N	205'430	15'543	23'401	14'295	17'115	34'238	59'562
N Ärzte	5‘455	481	1‘221	425	1‘020	871	2‘487
Koeffizient Logmodell	0.111*** (0.00)	0.059*** (0.00)	0.024 (0.27)	0.003 (0.80)	0.099*** (0.00)	-0.032 (0.21)	0.093*** (0.00)

Koeffizienten aus dem Gesamtmodell mit AGG, Franchisen, Spital-im-Vorjahr, und PCG, Zielvariable logarithmiert. Standardfehler in Klammern; Signifikanzniveaus: *** p<0.01, ** p<0.05, * p<0.1.

In der Tabelle sind für ausgesuchte Facharztgruppen die Koeffizienten des Indikators Spital-im-Vorjahr dargestellt. Der Indikator ist in vier Facharztgruppen statistisch signifikant positiv, in den anderen Gruppen hat er keinen signifikanten Einfluss.

Quelle: Daten der Sasis AG; Datenjahr 2015; eigene Berechnungen.

Über alle Facharztgruppen ist der Indikator Spital-im-Vorjahr in 13 Facharztgruppen signifikant positiv. In den restlichen Facharztgruppen ist er insignifikant. Ein signifikant negativer Effekt, welcher kontraintuitiv wäre, wurde in keiner Facharztgruppe festgestellt.

6.2.4 Indikator PCG

Auf die Bildung der pharmazeutischen Kostengruppen sind wir bereits in Abschnitt 5.2.2 eingegangen. Wir haben dabei eine PCG-Liste verwendet, welche für den Risikoausgleich in der Krankenversicherung entwickelt wurde (Trottmann et al. 2015). Die möglichen PCG sind in Tabelle 10 aufgelistet.

Tabelle 10 PCG, deren Einbezug getestet wurde

PCG	Label	PCG	Label
1	Asthma	13	Hormonsensitive Tumore
2	COPD/schweres Asthma	14	Krebs
3	Zystische Fibrose/Pankreasenzyme	15	Nierenerkrankungen
4	Hoher Cholesterinspiegel	16	Erkrankungen des Gehirns/Rückenmarks
5	Morbus Crohn und Colitis ulcerosa	17	Neuropathischer Schmerz
6	Depression	18	Parkinson
7	Diabetes Typ I	19	Psychose, Alzheimer und Sucht
8	Diabetes Typ II	20	Rheuma
9	Epilepsie	21	Erkrankungen der Schilddrüse
10	Glaukom	22	Transplantationen
11	Herzerkrankungen	23	Bluthochdruck
12	HIV/AIDS	24	ADHS

Für die Operationalisierung der PCG wurde eine Klassifikation getestet, welche im Zusammenhang des Risikoausgleichs zwischen den Krankenversicherern entwickelt wurde. Sie enthält 24 pharmazeutische Kostengruppen.

Quelle: Trottmann et al. (2015), Tabelle 4; Änderung bei PCG 8 und 23 (siehe Abschnitt 5.2.2).

Wie in Abschnitt 5.2.2 beschrieben, wurde eine PCG nur berücksichtigt, wenn über 30 Ärzte innerhalb der Facharztgruppe eine Mindestmenge an Medikamenten aus der entsprechenden PCG verschrieben haben. Die Mindestmenge entspricht dabei 1.8 DDD pro Praxis. Dies ist eine rein statistische Entscheidungsgrösse, eine inhaltliche Beurteilung, welche PCG für welche Facharztgruppen ins Modell gehören, haben wir nicht vorgenommen.

In Tabelle 11 ist gezeigt, welche PCG bei den jeweiligen Facharztgruppen ins Modell einbezogen wurden. Bei den Ärzten der Allgemein Inneren Medizin sind dies alle 24 PCG. Da es sich um eine grosse Facharztgruppe handelt, ist die Schwelle der 30 verschreibenden Ärzte schnell erreicht. Ähnliches gilt für die Ärzte der Psychiatrie und Psychotherapie, wo 13 der 24 PCG einbezogen werden. Spezialisierte Facharztgruppen wie beispielsweise die Ophthalmologie verschreiben hingegen nur aus einzelnen PCG Medikamente (zum Beispiel PCG 10, Glaukom). Die PCG 23 (Bluthochdruck), welche in der Bevölkerung eine sehr hohe Prävalenz aufweist, ist die PCG, welche am häufigsten ins Modell einbezogen wurde.

Tabelle 11 Einfluss der PCG, Gesamtmodell

	Allgemeine Innere	Chirurgie	Gynäkologie	Kardiologie	Kinder/ Jugend	Ophthal-mologie	Psychiatrie/ Psycho-therapie
N	205'430	15'543	23'401	14'295	17'115	34'238	59'562
N Ärzte	5'455	481	1'221	425	1'020	871	2'487
Berücksichtigte PCG	Alle 24	4, 6, 23	6, 13, 14, 21, 23	1, 4, 6, 8, 11, 21, 23	1, 24	10	1, 4, 6, 8, 9, 16, 17, 18, 19, 20, 21, 23, 24

Die PCG werden nur in das Modell einbezogen, wenn in einer Facharztgruppe mindestens 30 Ärzte eine Mindestmenge an Medikamenten aus dem Indikationsgebiet verschrieben haben (rein statistisches Kriterium, keine inhaltliche Beurteilung). Bei der Facharztgruppe Allgemeine Innere Medizin trifft dies auf fast alle PCG zu, bei anderen Facharztgruppen wie der Ophthalmologie nur vereinzelt (PCG 10, Glaukom).

Quelle: Daten der Sasis AG; Datenjahr 2015; eigene Berechnungen.

Die Koeffizienten der PCG sind im Excel Anhang «Polynomics_Wirtschaftlichkeitspruefungen_Regressionskoeffizienten_Stufe_1» dargestellt. In keiner der 30 Facharztgruppen gab es PCG, welche statistisch signifikant negative Koeffizienten erhielten. Einige Koeffizienten waren jedoch insignifikant negativ. Dies kam vor allem bei Gruppen mit geringen Wirkstoffmengen vor. Es wäre möglich, diese mit der Gruppe «keine Verschreibung» zu vereinen. Der PCG-Indikator würde dann erst ab einer gewissen Wirkstoffmenge berücksichtigt. Dieses Vorgehen entspricht zum Beispiel auch dem HCC-Modell, welches im US-amerikanischen Medicare-Programm zur Risikoadjustierung genutzt wird (Pope et al. 2000). Alternativ könnte die Wirkstoffmenge als kontinuierliche Variable spezifiziert werden (vgl. Abschnitt 5.2.2), um negative Koeffizienten ganz zu vermeiden.

6.3 Überblick über die erste Stufe

In Tabelle 12 sind Kennzahlen des Erklärungsgehalts im Modell mit hohen Franchisen, Spital-im-Vorjahr und PCG abgebildet. Zum Vergleich sind auch die Kennzahlen auch für eine Regression nur mit dem praxisspezifischen Effekt (PE) und mit einem Modell bestehend aus PE und AGG abgebildet. Das Modell mit nur einem praxisspezifischen Effekt entspricht dabei dem Vergleich der gewichteten Mittelwerte, wie er auch im RSS-Index praktiziert wird.

Neben dem R^2 und dem mittleren absoluten Prognosefehler (MAPE) sind auch die beiden Informationskriterien AIC (Akaike Information Criterion) und BIC (Bayesian Information Criterion) angegeben. Bei beiden Informationskriterien ist ein niedriger Wert besser (auch im negativen Bereich). Im Unterschied z. B. zum R^2 beurteilen diese Indikatoren den zusätzlichen Informationsgehalt von Einflussfaktoren. Sie werde schlechter, wenn zusätzliche Faktoren aufgenommen werden, welche wenig zum Erklärungsgehalt beitragen.

Bei allen Indikatoren schneidet das Modell mit der Morbiditätsinformation besser ab als Modelle mit lediglich einem Praxiseffekt oder mit AGG und Praxiseffekt. Besonders bei Ärzten, deren Patientenstamm bezüglich der erklärenden Variablen vom Durchschnitt abweicht, wird die Berücksichtigung der Morbiditätsvariablen zu einer deutlichen Verbesserung der Beurteilung führen.

Zwischen den Facharztgruppen gibt es deutliche Unterschiede. Beispielsweise ist die Verbesserung des R^2 in den Facharztgruppen Allgemein Innere Medizin, Kinder- und Jugendmedizin und Psychiatrie und Psychotherapie deutlich stärker als in den anderen Facharztgruppen. Der MAPE, der nicht so stark auf Ausreisser reagiert wie das R^2, verbessert sich stark bei den Facharztgruppen Allgemeine Innere Medizin, Gynäkologie, Kinder- und Jugendmedizin sowie der Ophthalmologie. Eine weitere Diskussion der Regressionsdiagnostik befindet sind in Abschnitt 11.3 im Anhang.

Tabelle 12 Erklärungsgehalt des Modells mit hohen Franchisen, Spital-im-Vorjahr und PCG

	Allgemeine Innere	Chirurgie	Gynäkologie	Kardiologie	Kinder/ Jugend	Ophthal-mologie	Psychiatrie/ Psycho-therapie
N	205'430	15'543	23'401	14'295	17'115	34'238	59'562
N Ärzte	5‘455	481	1‘221	425	1‘020	871	2‘487
Adj. R^2							
Volles Modell	0.877	0.742	0.899	0.766	0.768	0.865	0.526
PE + AGG	0.817	0.729	0.892	0.700	0.745	0.862	0.427
PE	0.284	0.704	0.633	0.576	0.339	0.580	0.333
MAPE							
Volles Modell	0.279	0.341	0.286	0.223	0.490	0.252	0.438
PE + AGG	0.330	0.347	0.292	0.246	0.508	0.259	0.478
PE	0.627	0.360	0.444	0.285	0.710	0.417	0.515
AIC							
Volles Modell	1‘158.8	6‘771.4	-12‘990.0	-9‘494.2	-5‘098.7	-1‘386.4	54‘506.4
PE + AGG	81‘834.9	7‘527.6	-11‘438.7	-5‘934.3	-3‘508.4	-644.7	65‘744.6
PE	362‘275.1	8‘877.1	17‘216.2	-1‘040.0	12‘778.3	37‘378.5	74‘740.8
BIC							
Volles Modell	3‘747.7	7‘146.3	-12‘482.2	-8‘956.9	-4‘703.6	-939.0	55‘468.9
PE + AGG	82‘254.5	7‘841.3	-11‘108.3	-5‘624.0	-3‘190.8	-298.7	66‘113.4
PE	362‘275.1	8‘877.1	17‘216.2	-1‘040.0	12‘778.3	37‘378.5	74‘740.8

Zielvariable logarithmiert. N: Stichprobengrösse; adj. R^2: korrigiertes Bestimmtheitsmass; MAPE: durchschnittlicher absoluter Prognosefehler; AIC: Akaikes Informationskriterium; BIC: Bayesianisches Informationskriterium; PE: Praxiseffekt; AGG: Alters- und Geschlechtsgruppen.

Die schwarz gedruckten Zahlen geben den Erklärungsgehalt des Modells mit Franchisestufen, Spital-im-Vorjahr und PCG an, die grauen Zahlen der Modelle nur mit dem Praxiseffekt (PE) bzw. mit dem PE und AGG. In allen Indikatoren schneidet das Modell mit der Morbiditätsinformation am besten ab. Bei Ärzten, deren Patientenstamm bzgl. der erklärenden Variablen vom Durchschnitt abweicht, sollte die Berücksichtigung der Morbiditätsvariablen zu einer deutlichen Verbesserung der Beurteilung führen.

Quelle: Daten der Sasis AG; Datenjahr 2015; eigene Berechnungen.

7 Praxisspezifischer Effekt und Indexberechnung

7.1 Verteilung des Punktschätzers für den praxisspezifischen Effekt

Die Berechnung des praxisspezifischen Effekts erfolgt grundsätzlich analog zur Berechnung der übrigen Koeffizienten. Allerdings wird der Effekt nicht im Vergleich zu einer Referenzkategorie betrachtet, sondern im Vergleich zum Durchschnitt der Facharztgruppe. So ist sichergestellt, dass der mittlere Praxiseffekt in allen Facharztgruppen null beträgt (siehe Spalte «Mittelwert» in Tabelle 13).

Die Verteilung der Praxiseffekte ist in Tabelle 13 dargestellt.[8] Im logarithmischen Modell entsprechen die Werte mal 100 ungefähr dem Praxiseffekt in Prozent. Ein ausgewiesener Praxiseffekt von 0.16 heisst also, dass die Praxis etwa 16 Prozent höhere Kosten hat, als man es aufgrund der Patientenstruktur erwarten würde. Die Zahlen im untransformierten Modell sind die Abweichungen vom Durchschnitt pro Praxis in Franken. Dieser Durchschnitt ist je nach Facharztgruppe unterschiedlich, daher sind die Zahlen nicht direkt vergleichbar. Ein solcher Vergleich findet sich bei den Resultaten zur Indexberechnung in Kapitel 7.3.

Auffällig ist der grosse Anstieg zwischen dem 95-Prozent-Perzentil, dem 99-Prozent-Perzentil und dem Maximum. Es gibt eine kleine Gruppe von Praxen, deren unerklärte Kosten pro Patient weit über den anderen Praxen liegen. Im untransformierten Modell ist der Unterschied besonders deutlich, was sicher als Nachteil dieses Modells zu sehen ist. In den meisten dieser Fälle ist zu erwarten, dass diese sehr hohen Werte durch ein spezielles Leistungsspektrum oder eine andere Praxisbesonderheit erklärbar sind. In einem eigenen Teilprojekt sollte im Anschluss an diese Studie geprüft werden, welches diese Besonderheiten sind und ob sie in das statistische Screening aufgenommen werden können.

Tabelle 13 Verteilung praxisspezifischer Effekt

	N	Mittelwert	Std. Abw.	p25	p50	p75	p95	P99	Max.	Schiefe
Logarithmiert	17'464	0.0	0.36	-0.16	0.00	0.17	0.49	0.96	5.83	-0.53
Untransformiert	17'464	0.0	476	-152	-38	83	432	1'290	162'239	64
Winsorisiert	17'464	0.0	294	-125	-22	95	387	830	23'018	10

Gewichtet nach Anzahl Patienten pro Praxis. Praxen aus Facharztgruppen mit unter 30 Praxen und Gruppenpraxen nicht dargestellt.

Die Praxiseffekte geben die durchschnittliche Differenz einer Arztpraxis zum Gesamtdurchschnitt der jeweiligen Facharztgruppe an. Ihr Mittelwert ist deshalb immer null. Die Werte im logarithmierten Modell zeigen (multipliziert mit 100) die Abweichungen in Prozent an, während die anderen beiden Modelle diese in Franken ausweisen.

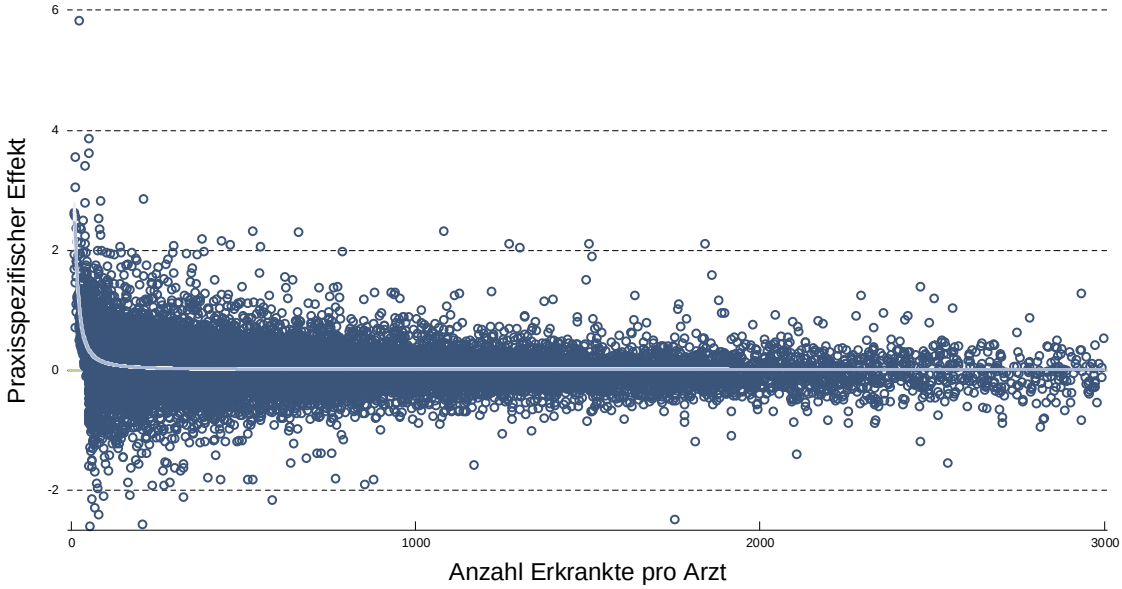
Quelle: Daten der Sasis AG; Datenjahr 2015; eigene Berechnungen.

In Abbildung 6 ist der Zusammenhang zwischen der Anzahl Erkrankter pro Arzt und dem berechneten, praxisspezifischen Effekt dargestellt. Die «Trichterform» der Punktewolke zeigt an, dass bei sehr kleinen Praxen (bis rund 200 Erkrankte) die Streuung des praxisspezifischen Effekts höher ist als bei grossen Praxen. Die hellblaue Trendlinie verläuft über weite Teile der

[8] In der Tabelle sind die Ergebnisse der Gruppenpraxen und der Facharztgruppen mit weniger als 30 Praxen nicht dargestellt. Diese sind sehr heterogene Gruppen, und bei der Interpretation der Praxiseffekte ist Vorsicht geboten.

Verteilung flach, es ist also kaum ein systematischer Zusammenhang zwischen der Praxisgrösse und dem praxisspezifischen Effekt auszumachen. Bei den allerkleinsten Praxen zeigt die Trendlinie plötzlich nach oben. Bei Praxen bis ca. 50 Erkrankten liegt also auch der berechnete Praxiseffekt systematisch höher. Dies ist entscheidend durch einen Selektionseffekt verursacht: Sehr kleine Praxen werden nur dann in das statistische Screening eingeschlossen, wenn sie über 100‘000 Franken an Kosten verursacht haben (siehe Abschnitt 11.2 im Anhang).

Abbildung 6 Praxisspezifischer Effekt und Anzahl Erkrankter



Jeder Punkt ist eine Arztpraxis, N = 17‘078, Arztpraxen mit über 3000 Patienten (N = 379) nicht dargestellt.

Die hellblaue Trendlinie verläuft sehr flach, über weite Teile der Verteilung ist also kaum ein Zusammenhang zwischen der Praxisgrösse und dem berechneten Praxiseffekt erkennbar. Bei den kleinen Praxen (unter etwa 200 Erkrankten) ist die Streuung der Punkte etwas ausgeprägter. Auffällig zeigt die Trendlinie bei den allerkleinsten Praxen (unter etwa 50 Erkrankten) nach oben. Dies ist der Stichprobenselektion geschuldet, denn sehr kleine Praxen werden nur dann eingeschlossen, wenn sie über CHF 100‘000 an Bruttokosten verursacht haben.

Quelle: Daten der Sasis AG; Datenjahr 2015; eigene Berechnungen.

7.2 Korrektur um Charakteristika des Praxisstandortes

Obwohl die Daten bereits um die unterschiedlichen Taxpunktwerte bereinigt wurden, ist es denkbar, dass die Behandlung von ähnlichen Patientenkollektiven je nach Region zu unterschiedlichen Kosten führt. Dies könnte z. B. durch kulturelle Einflüsse bedingt sein, oder durch den Einfluss der regional wichtigen medizinischen Ausbildungsstätten. Zudem können Faktoren wie der Urbanitätsgrad oder die Sozialhilfequote eine Rolle für die Kosten pro Patient spielen.

Wie in Abschnitt 4.1.2 beschrieben, korrigieren wir die praxisspezifischen Effekte in einer zweiten Regression um den Einfluss des Praxisstandortes (siehe auch Kaiser, 2016). In dieser zweiten Regression ist der praxisspezifische Effekt die Zielvariable (ein Wert pro Praxis). Die Regressionskoeffizienten werden als Durchschnitt über alle Facharztgruppen geschätzt, denn viele Facharztgruppen haben zu wenig Praxen, als dass eine Berechnung pro Facharztgruppe

zweckmässig wäre. Auch auf dieser Stufe wird eine Gewichtung mit der Anzahl Erkrankten pro Arzt vorgenommen.

7.2.1 Praxiskanton

Der Einfluss des Praxiskantons ist in der linken Hälfte von Tabelle 14 dargestellt.[9] Von den 25 Kantonen sind 17 nicht signifikant verschieden von der Referenzkategorie (Aargau). Die signifikanten Koeffizienten sind dunkelgelb (positiver Koeffizient) oder blau (negativer Koeffizient) markiert. Auffallend ist der stark negative Wert des Kantons Uri. Der Extremwert kommt wohl unter anderem durch die kleine Fallzahl an dort niedergelassenen Praxen zu Stande. Für eine stabile Berechnung der Koeffizienten sollte aus unserer Sicht eine Gruppe mindestens 30 Praxen aufweisen. Kantone mit weniger Praxen könnten in der Analyse mit strukturell ähnlichen Nachbarkantonen zusammengefasst werden.

Da viele Koeffizienten der Kantone insignifikant sind, stellt sich die Frage, ob das Modell nicht auch mit einer grösseren geografischen Einheit geschätzt werden könnte. Auf der rechten Seite von Tabelle 14 sind die Koeffizienten für die sieben Grossregionen des Bundesamts für Statistik dargestellt. Hier sind die Genferseeregion sowie Zürich signifikant positiv verschieden von der Referenzkategorie (Espace Mittelland). Die Höhe der Koeffizienten lässt sich nicht direkt mit den Koeffizienten auf Kantonsebene vergleichen, weil die Referenzgruppe nicht die gleiche ist.

Das R^2 ist in beiden Modellen sehr klein, die geografischen Variablen erklären also nicht viel der Varianz in den praxisspezifischen Effekten. Interessant ist der Vergleich der Informationskriterien AIC und BIC. Die beiden Informationskriterien «bestrafen» es, wenn zusätzliche Variablen mit wenig Erklärungsgehalt ins Modell aufgenommen werden. Das AIC ist leicht besser (weil geringer) im Modell nur mit den Grossregionen, das BIC ist leicht besser im Modell mit den Kantonen. Bei beiden Indikatoren liegen die Modelle aber sehr nahe beieinander. Es gibt keine Hinweise darauf, dass ein Modell besser ist als das andere.

Da die Kantonsindikatoren nicht viel zum Erklärungsgehalt des Modells beitragen, wäre es aus statistischer Sicht durchaus gerechtfertigt, die Kantone im Modell durch Grossregionen zu ersetzen. Der Vorteil liegt darin, dass stabilere Koeffizienten geschätzt werden und Ausreisserwerte wie der Kanton Uri nicht mehr auftreten. Da sich in kleinen Kantonen spezielle Konstellationen ergeben können, ist es möglicherweise fairer, die Praxen in der Grossregion zu vergleichen. Dies wäre jedoch eine Abkehr von dem bisherigen Vergleich innerhalb des Kantons. Die Frage muss auf der inhaltlichen Ebene weiter diskutiert werden.

[9] Gegenüber den Auswertungen der ersten Stufe ist die Datenbasis etwas eingeschränkt, weil die geografische Information nicht von allen Ärzten zur Verfügung stand. Bei den Auswertungen zum Kanton, welche aus Datenschutzgründen bei Sasis im Haus durchgeführt wurden, wurde zudem die Einschränkung gemacht, das nur Praxen mit über 50 Erkrankten und über CHF 100‘000 Leistungen eingeschlossen wurden. Um die Vergleichbarkeit zu gewährleisten, werden alle Auswertungen zu den geografischen Variablen auf dieser eingeschränkten Datenbasis vorgenommen

Tabelle 14 Koeffizienten des Kantons bzw. der Grossregion auf der zweiten Stufe

Kanton	Koeffizient	Grossregion	Koeffizient
AG	Referenz	Espace Mittelland	Referenz
AI	-0.027	Genferseeregion	0.051***
AR	-0.002	Nordwestschweiz	-0.010
BE	-0.005	Ostschweiz	-0.018
BL	-0.056**	Zentralschweiz	-0.014
BS	-0.017	Zürich	0.019*
FR	-0.039*		
GE	0.098***		
GL	-0.037		
GR	-0.024		
JU	-0.087*		
LU	-0.051**		
NE	-0.018		
NW	-0.083		
OW	-0.042		
SG	-0.057***		
SH	0.026		
SO	0.016		
SZ	0.024		
TG	0.001		
TI	0.023		
UR	-0.163**		
VD	0.018		
VS	0.003		
ZG	0.030		
ZH	0.010		
N	15'680		15'680
R^2	0.013		0.008
AIC	12'243.3		12'183.8
BIC	12'542.1		12'635.8

Zielvariable logarithmiert, Weitere erklärende Variablen: Facharztgruppe, Einwohnerdichte der Wohngemeinde; Signifikanzniveaus: *** p<0.01, ** p<0.05, * p<0.1.

17 von den 25 Kantonsindikatoren sind nicht signifikant von null verschieden. Beim Kanton Uri resultiert ein stark negativer Koeffizient, welcher wohl der kleinen Anzahl an Praxen geschuldet ist. Die Informationskriterien geben an, dass ein Modell mit nur den Grossregionen statistisch nicht schlechter ist als ein Modell mit den Kantonen.

Quelle: Daten der Sasis AG; Datenjahr 2015; eigene Berechnungen.

7.2.2 Weitere Charakteristika des Praxisstandortes

Es ist denkbar, dass Charakteristika des Praxisstandortes den Praxiseffekt beeinflussen. Insbesondere hatten wir die Hypothese aufgestellt, dass die Siedlungsdichte, die Sozialhilfequote oder der Ausländeranteil einen Einfluss haben können. In den Daten des Bundesamts für Statistik[10] stehen diese Informationen pro Gemeinde zur Verfügung. Aus diesen haben wir anhand der empirischen Verteilung 5 bis 6 Gruppen gebildet.

Wie in Tabelle 15 dargestellt, waren diese Indikatoren in allen Schätzungen insignifikant. Die Koeffizienten der höchsten Gruppen (höchste Einwohnerdichte, höchste Sozialhilfequote, höchster Ausländeranteil) haben die erwarteten positiven Vorzeichen, sie sind aber inhaltlich von limitierter Grösse (die Details sind in Abschnitt 11.4 im Anhang gezeigt). Der Einfluss einer hohen Sozialhilfequote ist dabei der stärkste: Praxen in einem Gebiet mit einer hohen Quote an Sozialhilfeempfängern haben im Durchschnitt einen um rund 4 Prozent höheren praxisspezifischen Effekt als andere Praxen. Da die Koeffizienten insignifikant und eher gering sind, ist es statistisch gerechtfertigt, sie nicht im Modell zu berücksichtigen. Die Auswirkungen einer Weglassung auf die Indexberechnung zeigen wir in Abschnitt 7.4.

Tabelle 15 Übersicht Einbezug regionale Charakteristika auf der zweiten Stufe

	Spez. 1 Kanton	Spez. 1 Grossregion	Spez. 2 Kanton	Spez. 2 Grossregion	Spez. 3 Kanton	Spez. 3 Grossregion
Region	Ref.: AG ▪ Pos. sign. GE ▪ Neg. sign. BL, BS, JU, LU, NE, SG, UR	Ref.: Espace Mittelland ▪ Pos. sign: Genfersee, Zürich	Ref.: AG ▪ Pos. sign. GE ▪ Neg. sign. BL, FR, JU, LU, SG, UR,	Ref.: Espace Mittelland ▪ Pos. sign: Genfersee, Zürich	Ref.: AG ▪ Pos. sign. GE ▪ Neg. sign. BL, FR, JU, LU, SG, UR	Ref.: Espace Mittelland ▪ Pos. sign. Genfersee, Zürich ▪ Neg. sign. Ostschweiz
Einwohnerdichte	5 Gruppen Alle insign.	5 Gruppen Alle insign.	5 Gruppen Alle insign.	5 Gruppen Alle insign.		
Sozialhilfequote	6 Gruppen Alle insign.	6 Gruppen Alle insign				
Ausländeranteil					5 Gruppen Alle insign.	5 Gruppen Alle insign.
N	15'680	15'680	15'680	15'680	15'678	15'678
Adj. R^2	0.014	0.009	0.013	0.008	0.011	0.005
AIC	12'167.3	12'217.3	12'183.8	12'243.3	12'201.9	12'282.5
BIC	12'657.5	12'554.3	12'635.8	12'542.1	12'653.9	12'581.2

Zielvariable logarithmiert, Weitere erklärende Variablen: Facharztgruppe.
Die berechneten Variablen Einwohnerdichte, Sozialhilfequote und Ausländeranteil hatten keinen signifikanten Einfluss auf den Praxiseffekt. Auch der Erklärungsgehalt ist kaum unterschiedlich,

Quelle: Daten der Sasis AG; Datenjahr 2015; eigene Berechnungen.

[10] Regionalporträts 2015: Gemeinden – Kennzahlen; Abgerufen auf https://www.bfs.admin.ch/bfs/de/home/statistiken/regionalstatistik/regionale-portraets-kennzahlen/gemeinden.gnpdetail.2016-0166.html.

Eine mögliche Erklärung dafür, dass die Gemeindeindikatoren statistisch nicht signifikant sind, liegt darin, dass sie zu unscharf sind. Die Bevölkerungsdichte beispielsweise kann auch in ländlicheren Gemeinden, welche eine kleine Fläche haben, hoch sein. Die Sozialhilfequote und der Ausländeranteil hingegen sind innerhalb der Gemeinden, besonders innerhalb der grossen Städte, oft stark unterschiedlich. Der Durchschnitt pro Gemeinde sagt über den Patientenstamm der Praxis daher nicht allzu viel aus.

Eine andere Einteilung der Schweizer Gemeinden ist die neungliedrige Klassifikation des Bundesamtes für Statistik. Sie berücksichtigt zur Klassifikation den Zentrumscharakter einer Gemeinde und die Hauptbeschäftigung der Einwohner. Die Koeffizienten für die neun Gemeindetypen sind in Tabelle 16 dargestellt. Die suburbanen und ländlichen Gemeinden haben gegenüber den Zentren geringere Kosten im Rahmen von rund 5 bis 8 Prozent. Die einkommensstarken Gemeinden weisen gegenüber den Zentren höhere Kosten aus.

Tabelle 16 Einbezug BFS-Gemeindetypen auf der zweiten Stufe (Logmodell)

	Spezifikation 4	Spezifikation 5
Spezifikation	6 Grossregionen; BFS Gemeindetypen;	26 Kantone; BFS Gemeindetypen;
Region	Vergleichsgruppe: Espace Mittelland ▪ Pos. sign.: Genfersee, Zürich ▪ Neg. sign. Ostschweiz	Vergleichsgruppe: AG ▪ Pos. sign. GE ▪ Neg. sign. BL, FR, JU, LU, NW, SG, UR
Zentren	Referenzgruppe	Referenzgruppe
Suburbane Gemeinden	-0.05*** (-6.59)	-0.05*** (-6.91)
Einkommensstarke Gemeinden	0.09*** (5.80)	0.09*** (5.54)
Periurbane Gemeinden	-0.03 (-1.58)	-0.02 (-1.54)
Touristische Gemeinden	0.01 (0.37)	0.02 (1.02)
Industrielle und tertiäre Gemeinden	-0.06*** (-4.80)	-0.05*** (-4.17)
Ländliche Pendlergemeinden	-0.08** (-3.20)	-0.08** (-3.14)
Agrar-gemischte Gemeinden	-0.07**(-3.19)	-0.06** (-2.74)
Agrarische Gemeinden	-0.06 (-0.68)	-0.052 (-0.60)
N	15'678	15'678
Adj. R^2	0.011	0.018
AIC	12'184.3	12'098.2
BIC	12'513.7	12'580.8

Standardfehler in Klammern; Signifikanzniveaus: *** p<0.01, ** p<0.05, * p<0.1.

Der Gemeindetyp nach der neungliedrigen Klassifizierung des Bundesamts für Statistik hat Einfluss auf die berechneten Praxiseffekte. So sind die Praxiseffekte in ländlichen Regionen durchschnittlich kleiner, in einkommensstarken Gemeinden durchschnittlich höher als in den Zentren. Es bleibt jedoch die inhaltliche Frage zu diskutieren, ob diese Unterschiede gerechtfertigt sind, oder z. B. auf ein grösseres Moral-Hazard-Problem in einkommensstarken Gemeinden hinweisen.

Quelle: Daten der Sasis AG; Datenjahr 2015; eigene Berechnungen.

Auch wenn die Effekte statistisch signifikant sind, bleibt die inhaltliche Frage zu diskutieren, ob eine Berücksichtigung zu einer faireren Beurteilung der Arztpraxen führt. Es ist beispielsweise

durchaus denkbar, dass in einkommensstarken Gemeinden, wo ein dichtes Angebot vorhanden ist, das Problem des Moral Hazard stärker ausgeprägt ist als in den ländlichen Regionen. In diesem Fall wäre es auch gerechtfertigt, dass die Indexwerte in ländlichen Regionen durchschnittlich geringer sind. Dies müsste dann nicht durch das Rechenmodell korrigiert werden.

7.3 Resultate der Indexberechnung

7.3.1 Indexberechnung mit dem Punktschätzer

Die Resultate der Indexberechnung wie durch Kaiser (2016) vorgeschlagen sind in Tabelle 17 dargestellt. Definitionsgemäss ist der gewichtete Durchschnitt (Gewichtung nach Anzahl Patienten) der Indexwerte in allen Facharztgruppen 100. Wie schon bei den praxisspezifischen Effekten ist die Verteilung der Indexwerte rechtsschief, die Spannweite der Werte über 100 ist also wesentlich grösser als die Spannweite der Werte unter 100.

Auffällig ist im logarithmierten und im untransformierten Modell der grosse Unterschied zwischen dem 95-Prozent-Perzentil, dem 99-Prozent-Perzentil und dem Maximum. Dabei nimmt das Maximum im logarithmierten und im untransformierten Modell sehr hohe Werte an. Diese sind mit hoher Wahrscheinlichkeit auf Praxisbesonderheiten zurückzuführen, welche im statistischen Screening nicht erkannt werden können. Es ist nicht plausibel, dass der Praxisstil alleine solche grosse Wirkung haben kann. Deutlich sichtbar ist beim Maximum der Effekt der Winsorisierung: Da in der Berechnung des praxisspezifischen Effektes die Ausreisser «gestutzt» wurden, ist auch der maximale berechnete Indexwert deutlich geringer.

Tabelle 17 Resultate der Indexberechnung

	N	Mittelwert	Std. Abw.	p50	p75	p95	p99	Max.	Schiefe
Logarithmiert	17'464	100	60	95	112	152	237	32'569	243.97
Untransformiert	17'464	100	44	95	112	153	246	4'790	7.45
Winsorisiert	17'464	100	28	98	114	148	183	613	0.65

Gewichtet nach Anzahl Patienten pro Arzt.

Die Indexwerte sind definitionsgemäss im Durchschnitt in jeder Facharztgruppe 100. Etwa 5 Prozent der Praxen haben einen Indexwert von über 150. Auffällig ist der grosse Unterschied zwischen dem 99%-Perzentil und dem Maximum. Im logarithmierten und im untransformierten Modell gibt es Ausreisser mit sehr hohen Indexwerten. Im winsorisierten Modell tritt dieses Problem aufgrund der Stutzung der höchsten Werte deutlich weniger auf.

Quelle: Daten der Sasis AG; Datenjahr 2015; eigene Berechnungen.

Mit der bisher angewandten Schwelle bei einem Indexwert von 130 werden durchschnittlich 18 Prozent der Praxen als solche mit einem statistisch auffälligen Kostenprofil identifiziert (siehe Tabelle 18). Die Überschneidungen zwischen den drei Berechnungsvarianten sind hoch, 15 Prozent der Praxen werden in allen drei Berechnungsvarianten als auffällig identifiziert.

In einigen Facharztgruppen (z. B. Chirurgie oder Ophthalmologie) werden im winsorisierten Modell etwas mehr Praxen als auffällig erkannt. Dies weist darauf hin, dass durch die Stutzung der Ausreisser bei der Modellberechnung Praxen identifiziert werden, welche hohe aber nicht «ausreisserisch hohe» Kosten pro Patient aufweisen. Ohne die Stutzung werden sie sozusagen von den Ausreissern «verdeckt».

In der Facharztgruppe der Psychiatrie und Psychotherapie werden mehr Praxen als auffällig identifiziert als in anderen Facharztgruppen. Da der Mittelwert des Indexes in allen Facharztgruppen 100 beträgt, muss dies an der Streuung liegen. Falls die Streuung vor allem *zwischen den Praxen* vorhanden ist, ist es legitim, dass dies im Index abgebildet wird. Es muss dann in einem zweiten Schritt abgeklärt werden, ob legitime Unterschiede (Praxisbesonderheiten) vorliegen, welche die Streuung erklären. Falls die Streuung jedoch *innerhalb der einzelnen Praxen* hoch ist, ist dies eher ein Hinweis auf Patienten mit sehr heterogenem Bedarf.[11] Auf diesen Punkt werden wir in Abschnitt 7.3.2 zurückkommen.

Tabelle 18 Anteil auffällige Praxen bei Indexwert 130

	Logarithmiertes Modell	Untransformiertes Modell	Winsorisiertes Modell	In allen Modellen
Alle Facharztgruppen	18%	18%	17%	15%
Allgemeine Innere	15%	14%	12%	11%
Chirurgie	17%	18%	20%	16%
Gynäkologie	17%	17%	14%	12%
Kardiologie	12%	13%	9%	8%
Kinder- und Jugendmedizin	14%	13%	13%	11%
Ophthalmologie	13%	14%	18%	11%
Psychiatrie und Psychotherapie	26%	27%	27%	24%

Der Anteil an Praxen mit auffälligen Kostenprofilen ist in allen drei Modellen ähnlich hoch. Welches Modell die höchste Anzahl an Praxen identifiziert, ist dabei pro Facharztgruppe unterschiedlich. 15 Prozent der Praxen werden in allen drei Modellen als auffällig identifiziert.

Quelle: Daten der Sasis AG; Datenjahr 2015; eigene Berechnungen.

In Tabelle 19 sind die Beobachtungen nach dem Perzentil der Kostenverteilung pro Facharztgruppe klassifiziert. In der Gruppe 90 – 100 Prozent befinden sich also alle Praxen, deren Kosten in den obersten 10 Prozent ihrer Facharztgruppe liegen. In einem Modell ohne Morbiditätskorrektur (bspw. in einem einfachen Mittelwertvergleich) werden diese Praxen fast ohne Ausnahme als auffällig markiert. Nach der Morbiditätskorrektur mittels logarithmierten Modells werden noch 83 Prozent der teuersten Praxen markiert, dafür werden mehr Praxen mit mittleren durchschnittlichen Kosten als auffällig markiert. Im winsorisierten Modell ist der beschriebene Effekt noch stärker, von den Praxen mit den höchsten Durchschnittskosten werden noch zwei Drittel als auffällig markiert.

[11] Die Unterscheidung von Streuung *zwischen* und Streuung *innerhalb* der Praxen wird auch in der Literatur gemacht. Eine Beurteilung von Arztpraxen wird dann als besonders zuverlässig angesehen, wenn die Streuung zwischen den Praxen im Vergleich zur Streuung innerhalb der Praxen hoch ist (siehe Adams et al., 2010b).

Tabelle 19 Anteil auffällige Praxen nach der Kostenverteilung

Perzentil der Kostenverteilung [1)]	Ohne Morbiditäts-korrektur	Logarithmiertes Modell	Untransformiertes Modell	Winsorisiertes Modell
0 – 10%	0%	0%	0%	0%
10 – 25%	0%	2%	2%	2%
25 – 50%	0%	4%	3%	3%
50 – 75%	4%	11%	11%	12%
75 – 90%	55%	38%	39%	41%
90 – 100%	98%	83%	83%	69%

[1)] Die Eingruppierung erfolgt nach der Kostenverteilung pro Facharztgruppe.

Ohne Morbiditätskorrektur – z. B. bei einem einfachen Mittelwertvergleich – werden fast alle Praxen mit den höchsten Kosten als auffällig markiert. Nach der Morbiditätskorrektur werden mehr Praxen in den unteren 75 Prozent der Kostenverteilung markiert, weil ihr Kostenprofil gegeben die gemessenen Morbiditätsindikatoren auffällig ist.

Quelle: Daten der Sasis AG; Datenjahr 2015; eigene Berechnungen.

Tabelle 20 gibt eine Übersicht über die Unterschiede zwischen dem winsorisierten und dem logarithmierten Modell. 95 Prozent aller Praxen werden in beiden Modellen gleich beurteilt. Die Gruppe, welche nur im logarithmierten Modell auffällig ist, ist mit 434 Praxen etwas grösser als die Gruppe, welche nur im winsorisierten Modell auffällig ist.

Tabelle 20 Vergleich der Auffälligkeiten im logarithmierten und im winsorisierten Modell

		Auffällig im winsorisierten Modell		Gesamt
		Nein	Ja	
Auffällig im logarithmierten Modell	Nein	14'006 (80%)	329 (2%)	14'335
	Ja	434 (3%)	2'695 (15%)	3'129
	Gesamt	14'440	3'024	

95 Prozent der Praxen werden mit dem logarithmierten und dem winsorisierten Modell gleich beurteilt. Die Zahl derer, die nur im logarithmierten Modell auffällig sind, ist etwas grösser als die, welche nur im winsorisierten Modell auffällig werden.

Quelle: Daten der Sasis AG; Datenjahr 2015; eigene Berechnungen.

Im Datensatz befinden sich viele Praxen mit einer kleinen Anzahl Erkrankter (z. B. bis 80 Erkrankte). Es ist zu erwarten, dass bei diesen Praxen eine höhere Streuung der Indexwerte beobachtet wird als bei den grösseren Praxen. In Tabelle 21 sind die Indexwerte dargestellt gruppiert nach Praxisgrösse. Zur Gruppierung verwenden wir die Quartile der Grössenverteilung nach Facharztgruppe. Die 25 Prozent kleinsten Praxen pro Facharztgruppe befinden sich also im ersten Quartil.

Auffällig ist, dass der durchschnittliche Indexwert der kleinen Praxen höher liegt als bei grösseren Praxen. Ab dem 2. Quartil der Grössenverteilung liegen die Indexwerte recht nahe beieinander. Es gibt mehrere Erklärungsmöglichkeiten für höhere Indexwerte bei kleineren Praxen. Ers-

tens gibt es bei kleinen Praxen einen Selektionseffekt: Praxen mit unter 50 Patienten werden nur dann ins Screening eingeschlossen, wenn sie über CHF 100‘000 Bruttoleistungen (also pro Patient über CHF 2‘000 im Jahr) verursacht haben. In den meisten Facharztgruppen, liegt dieser Wert deutlich über dem 95-Prozent-Perzentil der Verteilung (siehe Tabelle 6). Zweitens könnten Praxen mit wenigen Patienten versucht sein, mehr Leistungen für einzelne Patienten zu erbringen, während gut ausgebuchte Praxen dies weniger tun (müssen). Dann wäre der höhere Indexwert gerechtfertigt. Drittens könnten hohe Ausreisserwerte bei *einzelnen Patienten* in kleinen Praxen einen deutlich stärkeren Einfluss auf die Gesamtkosten haben als bei grossen Praxen. Dies sollte nicht im Indexwert abgebildet werden, denn es ist kein Praxiseffekt. Wir werden im folgenden Abschnitt und bei den Simulationsberechnungen am Ende von Abschnitt 8.3 auf diesen Punkt zurückkommen.

Tabelle 21 Indexwerte nach Praxisgrösse

Praxisgrösse	N	Mittel-wert	Std. Abw.	p50	p75	p95	p99
Logarithmiertes Modell							
1. Quartil pro Facharztgruppe	4‘337	121	172	107	137	219	417
2. Quartil pro Facharztgruppe	4‘377	105	47	99	118	168	251
3. Quartil pro Facharztgruppe	4‘371	99	34	96	110	143	217
4. Quartil pro Facharztgruppe	4‘379	96	40	92	108	143	210
Gesamt	**17‘464**	**100**	**60**	**95**	**112**	**152**	**237**
Winsorisiertes Modell							
1. Quartil pro Facharztgruppe	4'337	111	37	109	131	177	220
2. Quartil pro Facharztgruppe	4'377	104	28	103	119	153	185
3. Quartil pro Facharztgruppe	4'371	100	25	99	113	141	179
4. Quartil pro Facharztgruppe	4'379	97	28	96	111	143	176
Gesamt	**17'464**	**100**	**28**	**98**	**114**	**148**	**183**

Gewichtet mit Anzahl Patienten pro Arzt.

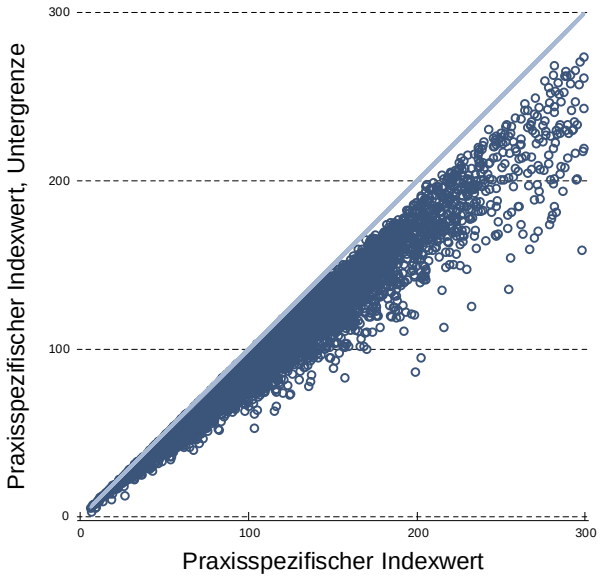
Praxen mit wenigen Erkrankten haben im Durchschnitt einen höheren Indexwert als Praxen mit vielen Erkrankten. Im winsorisierten Modell ist der Effekt nicht so deutlich.

Quelle: Daten der Sasis AG; Datenjahr 2015; eigene Berechnungen.

7.3.2 Indexberechnung mit Berücksichtigung des Vertrauensindikators

In Abbildung 7 sind der praxisspezifische Indexwert und die Untergrenze (nach der Berechnung in Abschnitt 4.4) dargestellt. Praxen, die nahe an der Mittellinie liegen, haben ein geringeres Streuungsmass. Falls ihre Kosten vom Mittelwert der Facharztgruppe abweichen (was sich in einem Indexwert ungleich 100 äussert), tun sie dies in allen AGG etwa gleich stark. Praxen mit hohem Streuungsmass liegen in Abbildung 7 unter der 45-Grad-Linie. Die Unsicherheit kommt daher, dass es starke Unterschiede zwischen den unterschiedlichen Beobachtungen (unterschiedlichen AGG) gibt. Es zeigt sich, dass Praxen mit einem hohen Indexwert tendenziell auch höhere Unsicherheitsindikatoren haben. Sie erfahren daher eher eine stärkere Veränderung, wenn der Index mit der Untergrenze berechnet wurde.

Abbildung 7 Untergrenze für den Indexwert



Jeder Punkt ist eine Arztpraxis, N = 17‘078, Arztpraxen mit über 3’000 Patienten (N = 386) nicht dargestellt.

Diese Abbildung zeigt den Zusammenhang des Indexwertes berechnet mit der Untergrenze (vertikale Achse) und berechnet mit dem Punktschätzer (horizontale Achse). Im Durchschnitt haben Praxen mit einem hohen Indexwert auch einen grösseren Unsicherheitsindikator als Praxen mit einem kleinen Wert.

Quelle: Daten der Sasis AG; Datenjahr 2015; eigene Berechnungen.

In der letzten Spalte in Tabelle 22 ist der Anteil an Praxen angegeben, bei welchen die Untergrenze des Indexwertes über 130 liegt. Gegenüber dem Punktschätzer wird der Anteil an auffälligen Ärzten deutlich reduziert, über alle Praxen sind noch 11 statt 18 Prozent auffällig. Am deutlichsten ist der Rückgang bei der Facharztgruppe der Psychiatrie und Psychotherapie, welche allerdings beim Punktschätzer einen hohen Anteil an auffälligen Praxen hatte. Dies ist ein Hinweis dafür, dass ein Teil des hohen Wertes dadurch zustande kam, weil durch die grosse Streuung *innerhalb* der Praxen der praxisspezifische Wert nicht eindeutig geschätzt werden konnte. Bei einer Berechnung des Indexes mit der Untergrenze sind in dieser Facharztgruppe denn auch nicht systematisch mehr Praxen auffällig als in anderen Facharztgruppen.

Tabelle 22 Anteil auffällige Praxen bei Verwendung der Untergrenze zur Beurteilung

	Logarithmiertes Modell	Logarithmiertes Modell, Indexberechnung mit Untergrenze
Alle Facharztgruppen	18%	11%
Allgemeine Innere	15%	9%
Chirurgie	17%	13%
Gynäkologie	17%	11%
Kardiologie	12%	8%
Kinder- und Jugendmedizin	14%	8%
Ophthalmologie	13%	11%
Psychiatrie und Psychotherapie	26%	12%

Die letzte Spalte gibt an, bei welchem Anteil der Praxen die Untergrenze des Indexwertes über 130 liegt. Der Anteil an auffälligen Praxen wird deutlich reduziert, über alle Praxen werden noch 11 statt 18% als auffällig identifiziert. Am stärksten ist der Rückgang in der Facharztgruppe der Psychiatrie und Psychotherapie, welche allerdings mit einem hohen Ausgangswert startete.

Quelle: Daten der Sasis AG; Datenjahr 2015; eigene Berechnungen.

Wie im vorigen Abschnitt gezeigt, ist der berechnete Indexwert für Praxen mit einer kleinen Anzahl Erkrankten systematisch höher als für grosse Praxen. Dies schlägt sich auch im Anteil an Praxen nieder, welcher als auffällig klassifiziert wird. In Tabelle 23 sind die Praxen nach der Anzahl Erkrankter in Gruppen eingeteilt. Es wird sichtbar, dass in der Gruppe mit den wenigsten Erkrankten deutlich am meisten Praxen als auffällig gelten. Bei einer Berechnung des Indexes mit dem Punktschätzer sind es beispielsweise 35 Prozent der kleinsten Praxen. Die höhere Anzahl an Auffälligen bleibt auch bei der Berechnung mit der Untergrenze bestehen. Dies ist ein Hinweis darauf, dass es nicht die grössere Streuung aufgrund der Zufallsschwankung ist, welche die kleinen Praxen häufige auffällig werden lässt, sondern systematisch höhere Werte (durch Stichprobenselektion oder höheren Moral Hazard). Bei der Simulation in Abschnitt 8.3 werden wir auf den Punkt mit der Zufallsschwankung zurückkommen (Tabelle 29).

Tabelle 23 Identifizierte Praxen mit Punktschätzer oder Untergrenze

Anzahl Erkrankte pro Praxis	N	Logarithmiert Punktschätzer	Logarithmiert Untergrenze
1. Quartil pro Facharztgruppe	4'337	35%	22%
2. Quartil pro Facharztgruppe	4'377	18%	11%
3. Quartil pro Facharztgruppe	4'371	11%	6%
4. Quartil pro Facharztgruppe	4'379	8%	5%

Bei den Praxen mit einer kleinen Anzahl an Erkrankten werden deutlich mehr Praxen auffällig klassifiziert als bei den grösseren Praxen. Dies bleibt auch bei der Verwendung der Untergrenze so. Im Abschnitt Simulation werden wir auf diesen Punkt zurückkommen.

Quelle: Daten der Sasis AG; Datenjahr 2015; eigene Berechnungen.

7.4 Spezifikationstests der zweiten Stufe

Wie in Kapitel 7.2 diskutiert, haben die Charakteristika des Praxisstandortes auf der zweiten Stufe kaum statistisch signifikanten Einfluss ausgeübt. In diesem Abschnitt prüfen wir ihren Einfluss auf die Indexberechnung. Dazu ist in Tabelle 24 dargestellt, wie sich der Anteil der als auffällig klassifizierten Praxen je nach Spezifikation der zweiten Stufe verändert.[12]

Als Referenzgrösse haben wir den Index nur nach der ersten Stufe berechnet (alle Aspekte der Indexberechnung bleiben gleich). So würden 17.6 Prozent der Praxen als auffällig identifiziert. Wird die Grossregion berücksichtigt, sinkt der Anteil auf 17.3 Prozent. Es handelt sich um 56 Praxen, die nicht mehr als auffällig gelten. Durch die Hinzunahme der Sozialhilfequote und der Facharztgruppe[13] sinkt die Anzahl identifizierter Praxen erneut um 37, so dass dann 17.0 Prozent der Praxen auffällig sind. Wird anstelle der Sozialhilfequote die Einwohnerdichte verwendet, sinkt der Anteil stärker und beträgt nun 16.9 Prozent. Werden beide Variablen Sozialhilfequote und Einwohnerdichte im Modell verwendet, sind ebenfalls noch 16.9 Prozent auffällig.

Tabelle 24 Anteile auffälliger Praxen bei unterschiedlicher Spezifikation der 2. Stufe

	N	Anzahl auffällige Praxen	Anteil auffällige Praxen
2. Stufe mit Facharztgruppe, Sozialhilfequote, Einwohnerdichte und Grossregion	15'766	2'661	16.9%
2. Stufe mit Facharztgruppe, Einwohnerdichte und Grossregion	15'766	2'660	16.9%
2. Stufe mit Facharztgruppe, Sozialhilfequote und Grossregion	15'766	2'682	17.0%
2. Stufe mit Facharztgruppe, Grossregion	15'766	2'719	17.3%
1. Stufe	15'766	2'775	17.6%

Durch die zusätzliche Bereinigung auf der zweiten Stufe werden 114 Praxen weniger als auffällig identifiziert (Gesamtmodell mit allen Variablen). Würden nur die Grossregionen verwendet, wären es 56 Praxen.

Quelle: Daten der Sasis AG; Datenjahr 2015; eigene Berechnungen.

[12] Die Datenbasis ist dabei gleich eingeschränkt wir in Kapitel 7.2 beschrieben.

[13] Der Mittelwert jeder Facharztgruppe beträgt null (siehe Abschnitt 7.1. Die Facharztgruppe muss nur dann ins Modell auf der zweiten Stufe genommen werden, wenn auch andere erklärende Variablen berücksichtigt sind.

8 Simulation zur Bestimmung der falsch Positiven und falsch Negativen

8.1 Fehler 1. und 2. Art

Die Treffgenauigkeit eines statistischen Modells lässt sich generell anhand der Fehler in den Voraussagen bestimmen. Dabei wird üblicherweise zwischen einem Fehler 1. Art und einem Fehler 2. Art unterschieden. Ein Fehler 1. Art (auch α-Fehler genannt) liegt vor, wenn bei einem Hypothesentest die zugrundeliegende Nullhypothese verworfen wird, obwohl sie in Wirklichkeit wahr ist. Die andere mögliche Fehlentscheidung bildet der Fehler 2. Art (auch β-Fehler genannt) ab. Davon spricht man, wenn man die Alternativhypothese fälschlicherweise zurückweist.

Im vorliegenden Fall der Wirtschaftlichkeitsprüfung ist die Nullhypothese, dass eine Arztpraxis effizient arbeitet und keine überhöhten Kosten aufweist. Die Alternativhypothese lautet dementsprechend, dass die Praxis ineffizient arbeitet und dadurch überhöhte Kosten aufweist. Daraus ergeben sich die die vier Möglichkeiten im Hypothesentest, die in Tabelle 25 abgebildet sind.

Tabelle 25 Entscheidungstabelle bei statistischen Hypothesentests

	Wahrer Sachverhalt: Nullhypothese (Praxis ist effizient)	Wahrer Sachverhalt: Alternativhypothese (Praxis ist ineffizient)
Statistisches Testergebnis negativ: Nicht-Verwerfen der Nullhypothese	Korrekte Entscheidung richtig negativ	Fehler 2. Art falsch negativ
Statistisches Testergebnis positiv: Verwerfen der Nullhypothese	Fehler 1. Art falsch positiv	Korrekte Entscheidung richtig positiv

Ein statistischer Hypothesentest kann jeweils vier Ergebnisse aufweisen. Entweder wird die Nullhypothese korrekterweise nicht verworfen oder korrekterweise zugunsten der Alternativhypothese verworfen. In diesem Fall spricht man von richtig negativen und richtig positiven Ergebnissen. Die Nullhypothese kann aber auch fälschlicherweise verworfen oder fälschlicherweise nicht verworfen werden. In diesem Fall spricht man von Fehlern der 1. und 2. Art bzw. von falsch positiven und falsch negativen Testergebnissen.

Quelle: Eigene Darstellung, Polynomics.

Ein statistisches Testergebnis kann jeweils negativ oder positiv ausfallen. Im ersten Fall führt dies üblicherweise zu einem Nicht-Verwerfen der Nullhypothese und im zweiten Fall zu einem Verwerfen der Nullhypothese zugunsten der Alternativhypothese. Je nachdem, ob die Null- oder Alternativhypothese wahr ist, kommt es neben den richtigen Entscheidungen zu falsch positiven Resultaten (Fehler 1. Art). Es werden also Praxen fälschlicherweise als ineffizient eingeschätzt, obwohl sie effizient sind. Auf der anderen Seite resultieren auch falsch negative Resultate (Fehler 2. Art), wo ineffiziente nicht als ineffizient erkannt werden.

Das Ziel ist eine hohe Treffgenauigkeit mit einer statistischen Methode zu erreichen und damit die Fehler 1. und 2. Art möglichst klein zu halten. Dies erhöht gleichzeitig die Wahrscheinlichkeit, korrekte Testergebnisse zu erhalten. Der Treffgenauigkeit sind jedoch Grenzen gesetzt, weil Massnahmen zur Reduzierung eines Fehlers 1. Art automatisch dazu führen, dass sich die Wahrscheinlichkeit für Fehler 2. Art erhöhen. Die beiden Fehler lassen sich also nicht gleichzeitig minimieren. Es besteht vielmehr ein Zielkonflikt und man muss abwägen, ob man eher falsch positive oder falsch negative Resultate verhindern möchte.

Die grosse Schwierigkeit im vorliegenden Fall ist, dass wir die α- und β-Fehler gar nicht bestimmen können, weil wir die Wahrheit nicht kennen. Wir haben keine Informationen darüber, ob die Praxen im Datensatz tatsächlich effizient oder ineffizient gearbeitet haben. Es besteht auch kein Test, welcher die (In-)effizienz eindeutig feststellen könnte (kein Referenzstandard). Die statistischen Modelle mit den echten Daten können dementsprechend nicht dahingehend überprüft werden, ob weniger oder mehr falsch positive und falsch negative Resultate auftreten. Die einzige Möglichkeit dafür besteht darin, dies anhand von Simulationen zu tun. Dabei simulieren wir alle Modelldaten selbst und generieren Praxen, welche ineffizient sind. Dadurch wird es möglich, die «Wahrheit» mit den statistischen Testergebnissen zu vergleichen und die Modelle auf ihre Treffgenauigkeit hin zu beurteilen.

8.2 Vorgehen Simulation

8.2.1 Berechnung erwartete Kosten

Um erwartete (effiziente) Kosten für jede Arztpraxis zu erhalten, wird das Modell der ersten Stufe (1) mit absoluten, das heisst nicht logarithmierten, Kosten y_{ij} geschätzt. Anschliessend werden für jede Praxis die erwarteten Kosten $\hat{y}_{ij}$ sowie die Residuen $\hat{\varepsilon}_{ij}$ ohne praxisspezifischen Effekt $\hat{a}_i$ aus den resultierenden Modellparameter berechnet. Die Formeln sind in den Gleichungen (10) und (11) aufgeführt. Die ökonometrische Schätzung sowie die anschliessende Berechnung der erwarteten Kosten und Residuen erfolgt für jede Facharztgruppe separat.

$$\hat{y}_{ij} = AGG_j\hat{\beta}_1 + X_{ij}\hat{\beta}_2 \tag{10}$$

$$\hat{\varepsilon}_{ij} = y_{ij} - \hat{y}_{ij} - \hat{a}_i \tag{11}$$

$$\text{für alle } i \in FAG_f \text{ und } f = \{1,2,\dots,F\}$$

8.2.2 Simulation Ineffizienz und Störterm

Die erwarteten Kosten $\hat{y}_{ij}$ stellen Kosten ohne Ineffizenz und Zufallsschwankung dar. In der Simulation werden in mehreren Durchläufen jeweils Ineffizienzwerte pro Praxis sowie Residuen pro Beobachtung generiert. Die Bildung der simulierten Kosten $\dot{y}_{ij}$ ist in Gleichung (12) dargestellt. Die Ineffizienzwerte I_i werden aus einer Exponentialverteilung gezogen und anschliessend multiplikativ auf die erwarteten Kosten $\hat{y}_{ij}$ aufgeschlagen.

$$\dot{y}_{ij} = \hat{y}_{ij} \times (1 + I_i) + \tilde{\varepsilon}_{ij} \text{ und } I_i \sim Exp(\lambda) \tag{12}$$

Ein Arzt gilt als ineffizient, wenn seine Kosten 30 Prozent über dem Mittelwert liegen. Die Parameter werden so gewählt, dass pro Durchlauf ca. 10 Prozent ineffiziente Praxen generiert werden.[14] Zur Generierung der Residuen $\tilde{\varepsilon}_{ij}$ verwenden wir die tatsächlichen Residuen $\hat{\varepsilon}_{ij}$ aus Gleichung (3). Die Residuen werden wie die zugehörigen erwarteten Kosten pro Facharztgruppe

[14] Mathematisch wurde eine Exponentialverteilung gesucht, für die zutrifft, dass 10 Prozent der Werte mindestens 30 Prozent über dem Durchschnitt liegen. Dies gilt für die Exponentialfunktion mit $\mu = 0.3$. μ entspricht dann der mittleren Effizienz pro Facharztgruppe. Ineffizient sind in der Simulation Ärzte, welche 30 Prozent über der mittleren Effizienz liegen. Für den Grenzwert muss also gelten: $\bar{\bar{y}}_i > 1.3 \times \bar{y} \rightarrow \bar{\bar{y}}_i > 1.3 \times 1.3 \times \bar{y} = 1.69 \times \bar{y}$, wobei $\bar{\bar{y}}$ die mittleren Kosten mit Ineffizienz und $\bar{y}$ die mittleren Kosten ohne Ineffizienz bezeichnen.

gebildet. Ziel ist, die Streuung in den Daten möglichst realitätsnah abzubilden. In jedem Simulationslauf wird pro Beobachtung ein Residuum $\tilde{\varepsilon}_{ij}$ aus der Verteilung der $\hat{\varepsilon}_{ij}$ gezogen und zu den erwarteten Kosten $\hat{y}_{ij}$ addiert.[15]

8.2.3 Überprüfung Treffsicherheit

Nach der Generierung der simulierten Kosten $\dot{y}_{ij}$ erfolgt die ökonometrische Schätzung des zweistufigen Modells sowie die Bildung der standardisierten Indexwerte zur Identifikation von ineffizienten Praxen. In jedem Durchlauf wird das zweistufige Modell dreimal geschätzt, dabei ist der Aufbau des Modells identisch, die Spezifikation der Zielvariablen jedoch unterschiedlich. Die simulierten Kosten $\dot{y}_{ij}$ werden einmal in absoluten Werten, einmal winsorisiert und einmal logarithmiert verwendet. Für jedes Modell wird die Treffsicherheit pro Durchlauf evaluiert. Dazu ermitteln wir die Anzahl der Praxen, die *richtig positiv*, *falsch positiv*, *falsch negativ* sowie *richtig negativ* klassifiziert werden (siehe Tabelle 26). Zusätzlich werden in jedem Durchlauf die Kennzahlen *Sensitivität*, *Spezifität* und *PPV* (positiver Vorhersagewert) berechnet.[16]

Tabelle 26 Kategorien zur Überprüfung der Treffsicherheit

	simulierter Praxiseffekt ≤ 30 Prozent	simulierter Praxiseffekt > 30 Prozent
Berechneter Indexwert ≤ 130	richtig negativ	falsch negativ
Berechneter Indexwert > 130	falsch positiv	richtig positiv

Die Ergebnisse werden in vier Kategorien unterteilt, um die Treffsicherheit zu untersuchen. Negativ bedeutet, dass der Praxiseffekt vom statistischen Screening als nicht auffällig identifiziert wurde, positiv hingegen, dass er als auffällig identifiziert wurde. Das Ziel ist es, die Anzahl richtig erkannter Arztpraxen (richtig positiv und richtig negativ) zu maximieren respektive die falsch erkannten Praxen (falsch positiv und falsch negativ) zu minimieren.

Quelle: Eigene Darstellung.

Die Formeln zur Berechnung der Kennzahlen sind in den Gleichungen (13) bis (15) aufgeführt. Die Kennzahl *Sensitivität* gibt Aufschluss darüber, wie hoch der Anteil an ineffizienten Arztpraxen ist, die richtig erkannt werden.

$$\text{Sensitivität} = 100 * \frac{\text{Anzahl richtig positiv}}{\text{Anzahl richtig positiv+Anzahl falsch negativ}} \tag{13}$$

Die Kennzahl *Spezifität* mach eine Aussage darüber, wie hoch der Anteil an effizienten Praxen ist, die richtig erkannt werden.

$$\text{Spezifität} = 100 * \frac{\text{Anzahl richtig negativ}}{\text{Anzahl richtig negativ+Anzahl falsch positiv}} \tag{14}$$

Die Kennzahl *PPV* (Positive Predictive Value) gibt an, wie hoch der Anteil an tatsächlich ineffizienten Arztpraxen unter den als ineffizient klassifizierten Praxen ist.

[15] Zur Beurteilung der Robustheit haben wir eine alternative Spezifikation getestet, in welcher die Ineffizienz additiv hinzugefügt wurde. Die Resultate waren ähnlich.

[16] Diese Kennzahlen sind etabliert, um die Testgüte zu beurteilen, z. B. bei medizinischen Diagnoseverfahren.

$$\text{PPV} = 100 * \frac{\text{Anzahl richtig positiv}}{\text{Anzahl richtig positiv+Anzahl falsch positiv}} \quad (15)$$

8.3 Ergebnisse der Simulation

Wir haben insgesamt 110 Simulationsdurchläufe durchgeführt, bei denen jeweils alle Kennzahlen berechnet wurden. Die Ergebnisse sind in Tabelle 27 zusammengefasst. In der Tabelle sind jeweils der Mittelwert und in Klammern die Standardabweichung der Kennzahlen *Sensitivität*, *Spezifität* und *PPV* angegeben. Bei der Interpretation der Ergebnisse ist es wichtig zu beachten, dass wir bei der Datengenerierung *keine Praxisbesonderheiten* eingebaut haben. Die generierten Arztpraxen (N = 17‘464) unterschieden sich nur in ihrer Patientenstruktur (systematische Komponente) und in der Zufallsschwankung. Die als falsch positiv identifizierten Praxen sind also nur deshalb positiv, weil ihnen durch die zufällige Ziehung hohe Residuen zugeordnet wurden.

Im linken (dunkelblauen) Teil der Tabelle sind die Resultate zur Indexberechnung mit dem Punktschätzer dargestellt (vgl. Absatz 7.3.1). Das Modell mit der logarithmierten Zielvariable (Logmodell) erreicht in allen drei Kennzahlen die besten Werte, und scheint daher am besten geeignet für das statistische Screening. Die Standardabweichungen der Kennzahlen sind gering, die Resultate der unterschiedlichen Simulationsdurchläufe schwanken also nicht sehr stark.

Im Logmodell beträgt die Sensitivität 90 Prozent, ein Zehntel der tatsächlich positiven Praxen bleiben also im Screening unerkannt. Von den richtig negativen, welche die Mehrheit stellen, werden 97 Prozent richtig erkannt. Für das statistische Screening zur Wirtschaftlichkeitsprüfung ist von besonderem Interesse der PPV, also der Anteil an positiv klassifizierten Praxen, der tatsächlich positiv ist. Dieser liegt im Logmodell bei 76 Prozent. Bei rund einem Viertel der Praxen, die als auffällig klassifiziert werden, ist die Klassifizierung also nicht korrekt. Das Modell mit der absoluten Zielvariable weist einen leicht geringeren PPV auf (72%). Deutlich schlechter ist das winsorisierte Modell: Nur 60 Prozent der positiv erkannten Praxen sind tatsächlich positiv.

Im rechten (grüngelben) Teil von Tabelle 27 sind die Kennzahlen bei einer Indexberechnung mit der Untergrenze dargestellt. Der PPV wird im Vergleich mit dem Punktschätzer stark erhöht auf gute 97 Prozent. Die Wahrscheinlichkeit, dass eine in Wahrheit negative Praxis nur durch die Zufallsschwankung als positiv klassifiziert wird, sinkt also auf 3 Prozent. Bei der Sensitivität zeigt sich jedoch der Kehrseite: Die Wahrscheinlichkeit, eine wirklich positive Praxis im Screening zu erkennen, sinkt auf unter 70 Prozent.

Der Vergleich der drei Rechenvarianten Logmodell, winsorisiert und untransformiert zeigt, dass der PPV im untransformierten Modell sogar noch etwas besser ist als im Logmodell. Dies liegt wohl daran, dass die Praxiseffekte im transformierten Modell eine hohe Streuung aufweisen. Im untransformierten Modell ist daher die Indexberechnung mit der Untergrenze besonders einschneidend. Die Sensitivität sinkt demnach auf niedrige 40 Prozent im untransformierten Modell. Insgesamt weist das Logmodell auch bei der Berechnung mit der Untergrenze eine bessere Treffergenauigkeit aus.

Tabelle 27 Kennzahlen pro Modell

	Indexberechnung mit Punktschätzer			Indexberechnung mit Untergrenze		
	Sensitivität	Spezifität	PPV	Sensitivität	Spezifität	PPV
M1 Logmodell	89.9% (0.9)	96.9% (0.2)	76.1% (1.2)	66.2% (1.5)	99.8% (0.05)	96.8% (0.5)
M2 Winsorisiert 95%-Perzentil	73.9% (3.1)	94.5% (1.0)	60.0% (3.2)	35.9% (3.3)	98.9% (0.5)	79.5% (4.9)
M3 Absolut	80.1% (1.5)	96.6% (0.2)	72.4% (1.1)	40.0% (1.4)	99.9% (0.01)	99.7% (0.2)

Für jede Kennzahl sind der Mittelwert und die Standardabweichung in Klammern angegeben. Die Kennzahlen sind ein Mass dafür, wie gut die drei Modelle darin sind, die Arztpraxen richtig zu klassifizieren. Das Logmodell schneidet bei jeder Kennzahl besser oder mindestens genauso gut ab, wie die beiden anderen Modelle.

Quelle: Simulierte Daten auf Basis der Daten der Sasis AG, eigene Berechnungen.

In Tabelle 28 sind die Kennzahlen Sensitivität, Spezifität und PPV für die Facharztgruppen mit den meisten Ärzten für das Logmodell ausgewiesen. Die Kennzahlen *Sensitivität* und *Spezifität* liegen nahe beieinander. Die *Kardiologen* haben bei Sensitivität (93%) und Spezifität (98%) die höchsten Werte. Die Facharztgruppe *Kinder- und Jugendmedizin* haben bei Sensitivität (84%) und Spezifität (92%) jeweils den niedrigsten Wert. Bei der Kennzahl *PPV* schwanken die Werte über die Facharztgruppen etwas stärker. Die Facharztgruppe *Kinder- und Jugendmedizin* hat mit 54 Prozent erneut den geringsten Wert.

Tabelle 28 Kennzahlen pro Facharztgruppe

	Indexberechnung mit Punktschätzer			Indexberechnung mit Untergrenze		
	Sensitivität	Spezifität	PPV	Sensitivität	Spezifität	PPV
Alle Facharztgruppen	89.9% (0.9)	96.9% (0.2)	76.1% (1.2)	66.2% (1.5)	99.8% (0.05)	96.8% (0.5)
Allgemeine Innere	89.4% (1.5)	97.7% (0.3)	81.1% (2.0)	72.1% (20.1)	99.9 (0.1)	96.1 (9.3)
Chirurgie	92.9% (4.7)	97.4% (0.9)	80.1% (6.5)	72.2% (7.6)	99.8 (0.2)	97.2 (2.9)
Gynäkologie	91.5% (3.2)	96.5% (0.8)	74.4% (5.2)	68.5% (5.3)	99.7 (0.2)	96.7 (2.4)
Kardiologie	92.4% (4.3)	98.3% (0.7)	85.6% (6.4)	74.9% (6.8)	99.9 (0.2)	98.3 (2.6)
Kinder- und Jugendmedizin	83.9% (4.1)	92.1% (1.0)	54.1% (5.0)	56.6% (5.1)	99.2 (0.3)	88.7 (4)
Ophthalmologie	93.2% (3.2)	97.9% (0.6)	83.1% (4.8)	71.9% (5.7)	99.8 (0.2)	98 (1.8)
Psychiatrie und Psychotherapie	91.0% (2.4)	96.3% (0.5)	73.4% (3.3)	63.2% (3.8)	99.8 (0.1)	97 (1.6)

Für jede Kennzahl sind der Mittelwert und die Standardabweichung in Klammern angegeben. Die Kennzahlen sind nach ausgewählten Facharztgruppen aufbereitet. Die Kennzahlen Sensitivität und Spezifität liegen verhältnismässig nah bei einander. Bei der Kennzahl PPV schwanken die Werte über die Facharztgruppen etwas stärker. Dies reduziert sich jedoch, wenn die Untergrenze verwendet wird.

Quelle: Simulierte Daten auf Basis der Daten der Sasis AG, Eigene Berechnungen.

Die geringere Testgüte bei der Facharztgruppe *Kinder- und Jugendmedizin* ist zum Teil dem Design der Simulation geschuldet: Die Praxen in dieser Facharztgruppe weisen eine starke

Konzentration in einzelnen Alters- und Geschlechtsgruppen auf. Wie in Gleichung (12) dargestellt, werden die tatsächlichen Residuen bei der Datengenerierung ohne Gewichtung mit der Anzahl Erkrankten gezogen. Nun sind die durchschnittlichen Kosten pro Patient sind bei dieser Facharztgruppe stark rechtsschief verteilt (siehe Tabelle 6), und die hohen Kostenfälle sind insbesondere den männlichen Patienten über 20 Jahre zuzuordnen. Da diese Patientengruppe im Verhältnis zu den anderen Alters- und Geschlechtsgruppen nur über eine geringe Anzahl an Erkrankten verfügt, ist es nicht realistisch, dass die Residuen mit einer gleichen Wahrscheinlichkeit gezogen werden wie alle anderen Gruppen. Dies führt dazu, dass es viele Praxen mit simulierten Residuen gibt, deren Durchschnitt grösser null ist und damit die Identifikation von tatsächlich ineffizienten Ärzten erschwert wird.

Anhand der Simulationsergebnisse lässt sich zusätzlich analysieren, inwiefern der höhere Anteil an Positivwerten unter den sehr kleinen Praxen auf die Zufallsschwankung zurückzuführen ist. Dies könnte zum Beispiel der Fall sein, wenn hohe Ausreisserwerte bei kleinen Praxen den praxisspezifischen Wert ungerechtfertigt verzerren. In Tabelle 29 ist die Testgüte nach Praxisgrösse dargestellt (Indexberechnung mit dem Punktschätzer). Dabei zeigt sich, dass die kleinen Praxen nicht systematisch mehr falsch Positive haben als andere Praxen. Es gibt also keinen Hinweis darauf, dass die höhere Anzahl Positivwerte auf hohe Ausreisserwerte zurückzuführen ist. Die Gründe dafür sind eher bei den beschriebenen Effekten der Selektion oder der Anreize zu suchen.

Tabelle 29 Testgüte nach Praxisgrösse

Anzahl Erkrankte pro Praxis	Sensitivität	Spezifität	PPV
1. Quartil pro Facharztgruppe	84.9% (1.9)	96.9% (0.3)	75.6% (2.1)
2. Quartil pro Facharztgruppe	90.3% (1.5)	96.9% (0.3)	76.4% (1.9)
3. Quartil pro Facharztgruppe	91.7% (1.4)	96.8% (0.3)	76.4% (2.2)
4. Quartil pro Facharztgruppe	92.7% (1.3)	96.7% (0.3)	75.9% (2.1)

Die Testgüte ist bei den kleinen Praxen etwa gleich gut wie bei den grösseren. Es gibt also keine Hinweise darauf, dass die höhere Anzahl an Positivwerten auf hohe Ausreisserwerte zurückzuführen ist.

Quelle: Simulierte Daten auf Basis der Daten der Sasis AG, Eigene Berechnungen.

9 Indexberechnung mittels Individualdaten

Zusätzlich zu den aggregierten wurde uns ein Datensatz mit individuellen Daten von Patienten (Individualdaten) zur Verfügung gestellt. Mit diesen Daten soll untersucht werden, ob ein Screening auf der Basis von Individualdaten oder eine Bereinigung von einzelnen hohen Patientenkosten vor der Aggregation die Treffsicherheit verbessert. Dies haben wir wiederum mit Hilfe einer Simulation untersucht.

9.1 Datengrundlage

Die Individualdaten enthalten Daten aus den Kantonen Bern und Aargau von drei Versicherern (gemeinsamer Marktanteil ca. 30%). Sie enthalten die direkten Kosten, die veranlassten Kosten und den Patientenrecord (eine detaillierte Beschreibung der Datenaufbereitung ist im Anhang enthalten). Der Datensatz enthält die folgenden sieben Facharztgruppen (siehe Tabelle 30):

- Allgemeine Innere Medizin
- Chirurgie
- Gynäkologie
- Kardiologie
- Kinder- und Jugendmedizin
- Ophthalmologie
- Psychiatrie und Psychotherapie

Als Morbiditätsindikatoren stehen wie im aggregierten Datensatz die Variablen «hohe Franchise», «Spital-im-Vorjahr» und «PCG» zur Verfügung. Für die PCG ist die vom Arzt an einen bestimmten Patienten verschriebene DDD-Menge erfasst (pro Arzt und Patient). Die DDD-Menge wurde einerseits pro Patient summiert. Dies dient als Morbiditätsindikator, da ein Patient in einer bestimmten PCG auch bei anderen Ärzten höhere Kosten aufweisen kann und nicht nur beim Arzt, der die Medikamente verschrieben hat. Andererseits wurde die DDD-Menge auch auf Ebene des Arztes über die AGG summiert und durch die Anzahl Patienten geteilt. Damit steht für jeden Arzt seine gesamte durchschnittliche pro Patient verschriebene DDD-Menge pro AGG zur Verfügung.

Tabelle 30 Übersicht Individualdatensatz

Facharztgruppe	Beobachtungen	Anzahl Patienten	Anzahl Praxen
Allgemeine Innere Medizin	266'818	240'470	701
Chirurgie	11'677	11'491	57
Gynäkologie	70'204	67'876	148
Kardiologie	17'478	17'033	54
Kinder- und Jugendmedizin	61'566	52'926	94
Ophthalmologie	72'056	67'817	84
Psychiatrie und Psychotherapie	17'014	16'508	282
Total	516'813	369'894	1'420

Insgesamt sind im Individualdatensatz 1'420 Praxen aus sieben Facharztgruppen enthalten. Die Anzahl Patienten liegt bei 369'894. Da ein Patient bei unterschiedlichen Ärzten in Behandlung sein kann, liegt die Anzahl Beobachtungen insgesamt mit 516'813 etwas höher.

Quelle: Abrechnungsdaten von drei Krankenversicherern; eigene Berechnungen.

9.2 Praxisspezifischer Effekt und Indexberechnung mit Individualdaten

9.2.1 Vergleich von Modellen mit Individualdaten

In einem ersten Schritt haben wir Modelle mit verschiedenen erklärenden Variablen auf den Individualdaten getestet. Die Zielvariable bilden in jedem Modell die logarithmierten Kosten. Tabelle 31 fasst die Ergebnisse zusammen. Zum einen wurde ein Modell nur mit den AGG verwendet (M1), das zweite Modell enthält zusätzlich zu den AGG die Morbiditätsindikatoren hohe Franchise, Spital-im-Vorjahr sowie die PCG als DDD-Menge pro Patient (M2). Im dritten Modell wurde anstatt der DDD-Menge pro Patient die DDD-Menge pro Patient und Arzt verwendet (M3).

Die Resultate zeigen, dass sich das Modell für die Facharztgruppen verbessert, wenn Morbiditätsindikatoren berücksichtigt werden (steigendes korrigiertes Bestimmtheitsmass adj. R^2 sowie sinkende Informationskriterien AIC bzw. BIC). Zudem schneidet das Modell mit PCG auf Arzt- und Patientenebene (M3) besser ab als das Modell mit PCG auf Patientenebene (M2). Für die folgenden Auswertungen wurde aus diesem Grund immer das Modell mit AGG, Spital-im-Vorjahr, Franchise hoch sowie PCG mit DDD-Menge auf Arzt- und Patientenebene verwendet.

Tabelle 31 Erklärungsgehalt untersuchter Modelle mit Individualdaten

	Allgemeine Innere	Chirurgie	Gynäkologie	Kardiologie	Kinder/ Jugend	Ophthal-mologie	Psychiatrie/ Psycho-therapie
N	266'203	11'640	70'106	17'457	61'433	72'000	16'948
Adj. R2: Korrigiertes Bestimmtheitsmass							
▪ M1	0.24	0.16	0.10	0.14	0.12	0.31	0.08
▪ M2	0.34	0.16	0.12	0.16	0.15	0.39	0.21
▪ M3	0.39	0.20	0.13	0.24	0.17	0.40	0.25
AIC: Akaikes Informationskriterium							
▪ M1	820'961	30'310	180'086	34'954	167'270	161'231	50'108
▪ M2	782'667	30'248	178'697	34'728	164'719	152'797	47'696
▪ M3	762'225	29'777	177'901	32'830	163'511	151'465	46'775
BIC: Bayesianisches Informationskriterium							
▪ M1	821'370	30'597	180'416	35'257	167'567	161'589	50'402
▪ M2	783'349	30'660	179'255	35'139	165'233	153'394	48'183
▪ M3	762'907	30'189	178'405	33'241	163'990	151'961	47'239

- M1 = nur AGG
- M2 = AGG, Spital-im-Vorjahr, Franchise hoch, PCG Patientenebene
- M3 = AGG, Spital-im-Vorjahr, Franchise hoch, PCG pro Patient und Arzt

Von den drei untersuchten Modellen mit logarithmierter Zielvariable schneidet bei allen Facharztgruppen das Modell M3 mit AGG, Spital-im-Vorjahr, Franchise hoch sowie PCG mit DDD-Menge auf Patient- und Arztebene am besten ab. Bei diesem Modell resultieren die höchsten Bestimmtheitsmasse (adj. R2) sowie die niedrigsten Werte für die beiden Informationskriterien AIC und BIC.

Quelle: Abrechnungsdaten von drei Krankenversicherern; eigene Berechnungen.

9.2.2 Vergleich von Modellen mit Individualdaten und aggregierte Daten

Um zu untersuchen, ob eine Indexbildung auf Basis von Individualdaten zu besseren Ergebnissen führt, wurden die Individualdaten auf Arzt- und AGG-Ebene aggregiert und die Resultate mit den Individualdaten verglichen. Dabei wurden die erklärenden Variablen in Tabelle 32 verwendet.

Tabelle 32 Vergleich der Individualdaten mit den aggregierten Daten

Individualdaten	Aggregierte Daten
▪ AGG ▪ Hohe Franchise (0/1) ▪ Spitalaufenthalt im Vorjahr (0/1) ▪ PCG mit DDD-Menge pro Arzt und Patient	▪ AGG ▪ Anteil Patienten mit hoher Franchise pro AGG und Arzt ▪ Anteil Patienten mit Spitalaufenthalt im Vorjahr pro AGG und Arzt ▪ PCG mit durchschnittliche DDD-Menge pro Patient pro AGG und Arzt

Quelle: Eigene Darstellung.

Der Anteil als auffällig identifizierte Praxen ist beim Modell mit Individualdaten mit 14 Prozent höher als beim Modell mit den aggregierten Daten mit gut 10 Prozent (Tabelle 33). Damit werden im aggregierten Modell rund 50 Praxen weniger als auffällig identifiziert. Wie auch schon in Abschnitt 7.3 beobachtet, sind die Indexwerte von Praxen mit einer geringen Anzahl Patienten (1. Quartil pro Facharztgruppe) deutlich höher als bei den übrigen Praxen. Der Anteil der als

auffällig klassifizierten Praxen ist in dieser Kategorie rund doppelt so hoch wie über alle Gruppen. Dieser Effekt ist bei allen Facharztgruppen zu beobachten. Eine Ausnahme bildet die Chirurgie, hier ist der Anteil in der Gruppe mit den grössten Praxen höher (nicht dargestellt in Tabelle 33).

Tabelle 33 Indexwert und Anteil auffällige Praxen nach Praxisgrösse für Individualdaten und aggregierte Daten

	1. Quartil	2. Quartil	3. Quartil	4. Quartil	Gesamt
Anteil auffällige Praxen (Indexwert über 130)					
▪ Individualdaten	29.9%	8.7%	8.2%	9.4%	14.1%
▪ aggregierte Daten	21.8%	5.6%	5.1%	10.2%	10.7%
▪ in beiden Modellen	20.9%	4.5%	4.8%	7.1%	9.37%
Anteil auffällige Praxen mit Untergrenze (Indexwert über 130)					
▪ Individualdaten	15.9%	1.4%	2.5%	3.4%	5.8%
▪ aggregierte Daten	5.6%	0.8%	1.4%	3.4%	2.8%
▪ in beiden Modellen	5.3%	0.6%	1.4%	2.0%	2.3%
Durchschnittlicher Indexwert (gewichtet mit Anzahl Patienten)					
▪ Individualdaten	112	97	102	98	100
▪ aggregierte Daten	105	94	101	101	100
Durchschnittliche Anzahl Patienten					
▪ Allgemeine Innere Medizin	149	282	405	684	380
▪ Chirurgie	90	163	236	336	204
▪ Gynäkologie	233	370	493	799	473
▪ Kardiologie	118	239	359	590	323
▪ Kinder- und Jugendmedizin	253	480	741	1174	654
▪ Ophthalmologie	347	673	939	1469	857
▪ Psychiatrie und Psychotherapie	20	36	56	131	60

Der Anteil auffälliger Praxen ist bei Individualdaten mit 14% höher als bei aggregierten Daten mit gut 10%. In beiden Modellen schneiden Praxen mit wenig Patienten (1. Quartil pro Facharztgruppe) deutlich schlechter ab. Der Anteil auffälliger Praxen sinkt stark, wenn man die Untergrenze des Indexwertes verwendet. Dieser Effekt ist bei den aggregierten Daten stärker.

Quelle: Abrechnungsdaten von drei Krankenversicherern; eigene Berechnungen.

Wenn wir die Untergrenze des Indexwertes verwenden (siehe Abschnitt 4.4), sinkt der Anteil der auffälligen Praxen in allen Gruppen stark. Dieser Effekt ist sogar stärker bei den aggregierten Daten (Abnahme von 10.7 auf 2.8%) als bei den Individualdaten (14.1 auf 5.8%). Im Vergleich mit Kapitel 7.3 fällt auf, dass die Berücksichtigung der Untergrenze bei den aggregierten Versichererdaten einen grösseren Einfluss hat als bei den Sasisdaten. Der Grund liegt wahrscheinlich in der deutlich geringeren Abdeckung: Es werden in den Versichererdaten weniger Patienten pro Praxis beobachtet. Damit ist die Berechnung des praxisspezifischen Effektes ungenauer.

9.3 Simulation

Die Simulation wurde analog dem Vorgehen in Kapitel 8 durchgeführt und anhand den dort beschriebenen Kennzahlen ausgewertet. Die Datengenerierung basiert dabei auf den absoluten Individualdaten. Für die Indexberechnung wurden Modelle mit unterschiedlicher Datenbasis berechnet:

- alle Beobachtungen
- Winsorisierung der Kosten über dem 95-Prozent-Perzentil pro Facharztgruppe
- Ausschluss von Beobachtungen mit Kosten über dem 95-Prozent-Perzentil pro Facharztgruppe

Für die Winsorisierung sowie den Ausschluss von Beobachtungen haben wir jeweils das 95-Prozent-Perzentil pro Facharztgruppe aus den Vorjahresdaten verwendet.

Zusätzlich zu den Individualdaten wurden die Kosten nach Bereinigung auf Ebene Arzt und AGG aggregiert. Sowohl bei den Individualdaten als auch bei den aggregierten Daten wurden nach der Bereinigung respektive Aggregation die Kosten logarithmiert.[17]

9.3.1 Ergebnisse der Simulation

Die Resultate der Simulation zeigen, dass bei Verwendung eines Modells mit Individualdaten (I1) die Anzahl der falsch Negativen sehr gering ist (0.2%, Tabelle 34). Dies führt zu einem hohen Wert bei der Kennzahl Sensitivität (welcher Anteil der Ineffizienten wird richtig erkannt). Ebenfalls einen hohen Wert zeigt die Kennzahl Spezifität auf (welcher Anteil der Effizienten wird richtig erkannt), obwohl der Anteil der falsch positiven Werte mit 3.1 Prozent höher ist (Praxen die als auffällig identifiziert werden obwohl sie es nicht sind). Da diese jedoch ins Verhältnis mit den richtig Negativen gesetzt werden, welche im Idealfall 90 Prozent ausmachen,[18] fällt dieser Wert weniger stark in Gewicht. Die Kennzahl PPV liegt deutlich niedriger, das heisst 76 Prozent der Auffälligen sind tatsächlich ineffizient. Dieser Wert kann durch die Winsorisierung nur geringfügig angehoben werden (I2). Werden Beobachtungen mit hohen Kosten hingegen ausgeschlossen, dann sinkt der Wert sogar auf 73 Prozent. Bei den aggregierten Daten (A1) zeigt sich ein anderes Bild. Die falsch negativen Werte sind höher (2.1%), hingegen liegen die falsch Positiven niedriger (0.7%). Dies widerspiegelt sich in den Kennzahlen. So liegt die Kennzahl PPV mit 92 Prozent deutlich höher als bei den Individualdaten. Eine Winsorisierung oder ein Ausschluss von hohen Kosten reduziert zwar die falsch Positiven, erhöht jedoch die falsch Negativen.[19]

Die Resultate der Simulation zeigen, dass die Individualdaten bei der Identifizierung der ineffizienten Ärzte nicht besser abschneiden. Insgesamt werden beim Modell mit den Individualdaten (I1) 48 Praxen falsch eingeteilt (falsch positiv oder falsch negativ), bei den aggregierten Daten

---

[17] Die Ergebnisse aus der Regression mit Individualdaten würden den Resultaten einer gewichteten Regression mit aggregierten Daten entsprechen, wenn die erklärenden Variablen innerhalb der Gruppen nicht variieren. Ein Modell geschätzt nur mit AGG und absoluten Kosten liefert somit dieselben Koeffizienten. Da die Kosten jedoch erst nach der Aggregation logarithmiert werden, entspricht der logarithmierte Mittelwert pro Gruppe nicht mehr dem Mittelwert der logarithmierten Kosten.

[18] Mit der Simulation wurden rund 10% der Ärzte Kosten zugewiesen, die 30% über den durchschnittlichen Kosten gemäss ihrer Patientenstruktur liegen.

[19] Zu beachten ist, dass hier die Winsorisierung auf den Originaldaten vorgenommen wurde. Die Werte sind also nicht direkt mit der Winsorisierung der Sasis-Daten zu vergleichen.

(A1) sind es mit 39 Praxen deutlich weniger. Eine Winsorisierung oder der Ausschluss von einzelnen Beobachtungen verbessert die Treffgenauigkeit nicht in allen Fällen.

Tabelle 34 Kennzahlen pro Modell

		falsch negativ	falsch positiv	Sensitivität	Spezifität	PPV
I1	Individualdaten	0.2%	3.1%	97.7% (1.5%)	96.5% (0.7%)	76.0% (4.8%)
I2	Individualdaten winsorisiert 95-P.	0.3%	2.7%	96.8% (1.8%)	97.0% (0.6%)	78.6% (4.5%)
I3	Individualdaten Ausschluss 95-P.	0.1%	3.8%	98.9% (1.0%)	95.8% (0.7%)	72.7% (4.6%)
A1	aggregierte Daten	2.1%	0.7%	79.8% (3.9%)	99.2% (0.3%)	92.2% (3.1%)
A2	aggregierte Daten winsorisiert 95-P.	2.8%	0.4%	72.7% (4.2%)	99.6% (0.2%)	95.0% (2.4%)
A3	aggregierte Daten Ausschluss 95-P.	2.9%	0.4%	71.8% (4.3%)	99.5% (0.2%)	94.2% (2.6%)

Für jede Kennzahl ist der Mittelwert von 250 Simulationen angegeben (Standardabweichung in Klammer). Die Kennzahlen sind ein Mass dafür, wie gut die Modelle darin sind, die Praxen richtig zu klassifizieren. Die Modelle mit den aggregierten Daten schneiden bei der Treffsicherheit etwas besser ab als die Modelle mit den Individualdaten.

Quelle: Abrechnungsdaten von drei Krankenversicherern; eigene Berechnungen.

Zwischen den einzelnen Facharztgruppen schwankt der PPV beim Modell mit den Individualdaten (I1) zwischen rund 70 (Psychiatrie) und 83 Prozent (Kardiologie). Aber auch innerhalb der Facharztgruppen schwanken die Werte relativ stark in den simulierten Durchläufen. Durch die teilweise geringe Anzahl Ärzte pro Facharztgruppe (siehe Tabelle 30) können jedoch einige anders klassifizierte Ärzte die Kennzahl schon stark beeinflussen. Bei den aggregierten Daten (A1) schwankt der PPV zwischen 83 (Psychiatrie) und 96 Prozent (Allgemeine Innere Medizin) je nach Facharztgruppe.

Weiter haben wir untersucht, wie sich die Kennzahlen nach Praxisgrösse unterscheiden. Es zeigt sich, dass bei aggregierten Daten der PPV für die kleinsten Praxen (1. Quartil) kaum niedriger ist als bei den übrigen Praxen. Dieser Effekt ist unabhängig vom gewählten Modell. Es gibt also bei den kleinsten Praxen nicht systematisch mehr falsch positive aufgrund der Zufallsschwankung als bei anderen Praxen.

Tabelle 35 PPV nach Praxisgrösse für Individualdaten und aggregierte Daten

PPV	1. Quartil	2. Quartil	3. Quartil	4. Quartil	total
Individualdaten (I1)	61.8%	65.0%	66.3%	67.8%	65.0%
aggregierte Daten (A1)	87.1%	88.5%	88.2%	90.3%	88.6%

Quelle: Abrechnungsdaten von drei Krankenversicherern; eigene Berechnungen.

10 Fazit

Das Ziel der vorliegenden Studie war die Weiterentwicklung der statistischen Verfahren, welche den ersten Teil der Wirtschaftlichkeitsprüfungen bei Arztpraxen bilden. Dabei standen fünf Aspekte im Vordergrund: Erstens sollte das bisherige mathematische Verfahren im Lichte der internationalen Fachliteratur diskutiert und beurteilt werden. Zweitens sollte eine bessere Berücksichtigung der Morbidität des Patientenstamms vorgeschlagen und empirisch implementiert werden. Drittens war zu prüfen, ob Charakteristika des Praxisstandorts ebenfalls deutlich zur Erklärung der Kosten pro Praxis beitragen. Viertens sollte die Treffgenauigkeit des Verfahrens beurteilt und Möglichkeiten aufgezeigt werden, wie die Anzahl an falsch positiv klassifizierten Praxen reduziert werden kann. Fünftens sollte beurteilt werden, ob eine Berechnung mit Daten auf der Ebene der individuellen Patienten wesentliche Verbesserungen der Treffergenauigkeit erreichen könnte. Die Hauptergebnisse zu den fünf Aspekten werden im Folgenden kurz zusammenfasst.

Grundsätzlich halten wir das bisher angewandte *zweistufige Verfahren* – bestehend aus einer Fixed-Effects-Schätzung zur Berechnung eines praxisspezifischen Effektes auf der ersten Stufe und einer Bereinigung dieses Effekts auf der zweiten Stufe – für eine gute Modellspezifikation. Insbesondere ist die Fixed-Effects-Schätzung dafür geeignet, den spezifischen Einfluss einer Praxis rechnerisch von Einflussfaktoren des Patientenstamms (Morbiditätsindikatoren) sowie von Streuung aufgrund der Zufallsschwankung zu trennen.

Die geprüften *Morbiditätsindikatoren* Franchisestufe, Spital-im-Vorjahr und pharmazeutische Kostengruppen (PCG) schätzen wir für den Einbezug ins Modell auf der ersten Stufe als geeignet ein. Sie haben bei der Mehrzahl der Facharztgruppen einen statistisch signifikanten Einfluss auf die Kosten, und verbessern die allgemeine Güte des Schätzmodells gegenüber einer Schätzung nur mit den Faktoren Alter und Geschlecht. Besonders bei Praxen, welche in Bezug auf diese Faktoren vom Durchschnitt abweichen, erwarten wir durch die Morbiditätskorrektur eine verbesserte Beurteilung.

Als *weitere Einflussfaktoren auf der zweiten Stufe* haben wir den Einbezug von *Charakteristika des Praxisstandorts* geprüft. Konkret hatten wir die Hypothese aufgestellt, dass die Sozialhilfequote, die Einwohnerdichte oder der Ausländeranteil den errechneten Praxiseffekt beeinflussen könnten. Diese Faktoren erwiesen sich jedoch in den Schätzungen als statistisch nicht signifikant und auch von limitierter Grösse, so dass sie nicht zwingend ins Modell einbezogen werden müssen. Ein möglicher Grund für den limitierten Einfluss ist, dass diese Indikatoren nur auf Gemeindeebene zur Verfügung stehen und dies zu ungenau ist. So haben beispielsweise grosse Städte typischerweise sowohl Quartiere mit sehr hoher als auch sehr geringer Sozialhilfequote. Der Durchschnitt über die ganze Gemeinde ist wenig aussagekräftig.

Die *Testgüte des statistischen Verfahrens* (Anzahl falsch positiver und Anzahl falsch negativer Praxen) haben wir anhand von Simulationsrechnungen beurteilt. Dabei standen zwei Themen im Vordergrund: Erstens testeten wir zwei Alternativen zur bislang verwendeten logarithmischen Transformation der Zielvariablen. Die Simulationsrechnungen ergaben jedoch, dass diese Varianten betreffend der Testgüte der logarithmischen Transformation unterlegen waren. Zweitens entwickelten wir einen «Unsicherheitsindikator» (ähnlich einer Standardabweichung) für den praxisspezifischen Effekt. Dieser kann dazu verwendet werden, einen «Vertrauensbereich» (ähnlich einem Konfidenzintervall) für den praxisspezifischen Effekt zu berechnen. Wird anstelle des Punktschätzers (Durchschnitt) die Untergrenze des Vertrauensbereichs zur Indexberechnung verwendet, sinkt die Anzahl auffällig identifizierter Praxen um rund die Hälfte. Unsere

Simulationsauswertungen ergaben zudem, dass die Anzahl falsch positiver Praxen damit deutlich reduziert werden kann. Dieser Rückgang hat jedoch einen Preis: Durch das strengere Kriterium steigt die Anzahl falsch negativer Praxen an. Es ist also eine Abwägung zu treffen zwischen dem Risiko, Praxen zu identifizieren, die korrekterweise nicht unwirtschaftlich sind (falsch positive) und dem Risiko, Praxen nicht zu erkennen, die eigentlich unwirtschaftlich wären (falsch negative).

Die *Berechnungen mit patientenindividuellen Daten* ergaben, dass diese Daten unter den gegebenen Umständen keine deutlich bessere Testgüte erreichten als die bislang verwendeten aggregierten Daten. Obwohl durch die Aggregation Information verloren geht, scheint es in der aktuellen Situation nicht zwingend angezeigt, auf eine Berechnung mit individuellen Daten umzustellen. Neu wäre die Situation zu beurteilen, wenn die Datenbasis erweitert werden könnte, zum Beispiel durch eine Vollerhebung der patientenindividuellen Daten, oder durch eine Ergänzung mit diagnosebezogenen Indikatoren, welche im Ausland häufig zur Beurteilung der Wirtschaftlichkeit von Arztpraxen eingesetzt werden.

11 Anhang

11.1 Transformation der Zielvariable

11.1.1 Problem der Rücktransformation

Kaiser (2016) führt aus, dass der funktionale Zusammenhang zwischen den Einflussfaktoren und der Zielvariable im logarithmierten Modell besser geschätzt werden kann. Der Nachteil ist jedoch, dass die berechneten Koeffizienten den Einfluss einer erklärenden Variablen auf den erwarteten Wert der *logarithmierten Zielvariable* angeben, nicht den Einfluss auf die Zielvariable selbst. Formal ausgedrückt schätzt das Modell:

$$\begin{gathered} ln(y) = x\beta + \epsilon \text{ mit } E(\epsilon|x) = 0 \\ E(\ln(y)\,|x) = x\beta \end{gathered} \tag{16}$$

Auf der logarithmierten Skala gilt, dass die Residuen im Mittelwert null und unabhängig von den erklärenden Variablen sind ($E(\epsilon|x) = 0$). Daraus folgt jedoch nicht, dass der Erwartungswert der rücktransformierten Residuen gleich eins und unabhängig von den erklärenden Variablen ist ($E(\exp(\epsilon)|x) \neq 1$). Er ist insbesondere dann abhängig von den erklärenden Variablen, wenn die Varianz der Residuen auf der logarithmischen Skala mit den erklärenden Variablen korreliert (Heteroskedastie). Liegt Heteroskedastie vor, entsprechen die geschätzten Beta-Koeffizienten nicht direkt den Semielastizitäten, sondern es müsste bei strenger Anwendung auch der Unterschied in den Residualvarianzen beachtet werden. Eine detaillierte Diskussion dieses Themas findet sich in (Manning, 1998; Manning und Mullahy, 2001).

Manning und Mullahy (2001) empfehlen, beim Vorliegen von Heteroskedastie eine Schätzung mit dem Verfahren der Generalized Linear Model (GLM). Das in Gleichung (1) beschriebene Fixed-Effects-Modell müsste in diesem Fall mit einer individuellen [0/1]-Variable pro Praxis spezifiziert werden, anstelle mit dem von Kaiser (2016) vorgeschlagenen Within-Schätzer. Bei grossen Facharztgruppen sind die Schätzmodelle sehr rechenintensiv und konvergieren in Einzelfällen nicht. Wir beurteilen dieses Verfahren daher als zu instabil für einen operativen Einsatz in den Wirtschaftlichkeitsprüfungen. Wir empfehlen, das logarithmierte Modell zu verwenden, wie es von Kaiser (2016) beschrieben wurde, idealerweise mit dem Unsicherheitsindikator gemäss Abschnitt 4.4.

11.1.2 Approximative Rücktransformation

Der in Kapitel 4.2 berechnete Indexwert $\widehat{U}_i$ ist eine relative Grösse. Approximativ gibt sie an, um wie viel Prozent die durchschnittlichen Patientenkosten pro Praxis i vom Durchschnitt aller Praxen abweichen. Soll daraus eine absolute Grösse berechnet werden, können die beobachteten durchschnittlichen Kosten einer Praxis mit dem Kehrwert des Indexes multipliziert werden (geteilt durch 100, wenn der Index in Prozent ausgedrückt ist). Eine Praxis mit einem Indexwert 145 und beobachteten Kosten von CHF 1‘000 hätte demnach erwartete Kosten von 1‘000 / 1.45 = 690, gegeben der durchschnittliche Zusammenhang zwischen Kosten und den erklärenden Variablen. Die praxisspezifischen Zusatzkosten, welche nicht durch den Patientenstamm erklärt werden können, betragen dementsprechend CHF 310. Die Formel in Gleichung (17) beschreibt, dieses Vorgehen formal:

$$\hat{u}_i^{CHF} = \bar{y}_i \times \left(1 - \frac{1}{\exp(\widehat{U}_i)}\right) \text{ und } \bar{y}_i = \frac{\sum y_{ij}}{\sum Anzahl\ Patienten_{ij}} \tag{17}$$

$$\hat{\bar{y}}_i = \bar{y}_i - \hat{u}_i^{CHF}$$

Um approximativ den praxisspezifischen Effekt in Franken $\hat{u}_i^{CHF}$ zu erhalten, werden die tatsächlichen durchschnittlichen Kosten pro Patient $\bar{y}_i$ mit dem Faktor $\left(1 - 1/\exp\left(\hat{U}_i\right)\right)$ multipliziert. Soll zusätzlich ein Schätzwert für die aus dem Modell prognostizierten durchschnittlichen Kosten pro Patient $\hat{\bar{y}}_i$ berechnet werden, muss von den tatsächlichen durchschnittlichen Kosten $\bar{y}_i$ der praxisspezifische Effekt $\hat{u}_i^{CHF}$ subtrahiert werden.

Ein ähnliches Vorgehen wird bei der Effizienzmessung von Unternehmen im regulatorischen Kontext häufig angewandt. Beispielsweise führt die deutsche Bundesnetzagentur einen Effizienzvergleich von Strom- und Gasnetzbetreibern durch. Aus einem statistischen Benchmarkingverfahren mit den Methoden «Stochastic Frontier Analysis» (SFA) und «Data Envelopment Analysis» (DEA) resultiert ein Effizienzwert in Prozent. Dieser Prozentwert wird verwendet, um Vorgaben zu machen, um welchen absoluten Betrag ein Unternehmen seine Preise senken muss. Dies geschieht durch Multiplikation des Effizienzwertes mit den beobachteten, gesamten Erlösen des Unternehmens (siehe Bundesministerium für Justiz und für Verbraucherschutz, Anreizregulierungsverordnung – ARegV).

11.1.3 Indexberechnung für die Zielvariable in Levels

Werden auf der ersten Stufe die Kosten in absoluten anstatt logarithmierten Werten verwendet, dann sind der praxisspezifische Effekt $\hat{a}_i$ sowie das Residuum der zweiten Stufe $\hat{u}_i$ als absolute Abweichung vom Durchschnitt und nicht als prozentuale Differenz zu interpretieren (siehe Gleichungen 1 bis 3). Um für dieses Modell einen standardisierten Index zu berechnen, haben wir zunächst die Kosten bestimmt, welche auf Basis des Modells für die Praxis zu erwarten sind (durchschnittliche Kosten bei gegebenem Patientenstamm ($\hat{\bar{y}}_i$)). Dazu haben wir von den tatsächlichen Durchschnittskosten pro Praxis $\bar{y}_i = \frac{1}{J}\sum y_{ij}$ das Residuum aus der zweiten Stufe $\hat{u}_i$ abgezogen:

$$\hat{\bar{y}}_i = \bar{y}_i - \hat{u}_i \tag{18}$$

Der nichtstandardisierte Indexwert wird berechnet als Verhältnis aus den tatsächlich beobachteten und den erwarteten Durchschnittskosten $(\bar{y}_i/\hat{\bar{y}}_i) \times 100$. Der Quotient kann als relative Grösse interpretiert werden und gibt an, um wieviel Prozent die Durchschnittskosten der Praxis über oder unter den Kosten liegen, welche die Praxis nach dem zweistufigen Modell bei gegebenem Patientenstamm haben sollte.

Analog dem logarithmierten Modell (siehe Kapitel 4.2) wird auch dieser Index mit dem Durchschnitt über alle Arztpraxen in einer Facharztgruppe standardisiert:

$$S_f = 100/\frac{1}{N}\sum_{i=1}^{N}\frac{\bar{y}_i}{\hat{\bar{y}}_i} \tag{19}$$

Somit können auch für dieses Modell standardisierte Indexwerte berechnet werden:

$$\hat{U}_i = S_f \times \frac{\bar{y}_i}{\hat{\bar{y}}_i} \text{ für alle } i \text{ und } f \in \{1,2 \ldots, F\}. \tag{20}$$

11.2 Daten aus dem Sasis Datenpool/Tarifpool und Aufbereitung

11.2.1 Übersicht über die Datensätze

Die Daten aus dem Datenpool der Sasis AG werden bereits heute zur Berechnung der Wirtschaftlichkeitsprüfungen eingesetzt. Es ist aktuell die einzige Datenquelle, wo die Gesamtheit der in der Schweiz zu Lasten der obligatorischen Krankenversicherung abgerechneten medizinischen Leistungen erfasst werden. Diese Daten bilden die Grundlage für die in Kapitel 6 bis 8 vorgestellten Berechnungen. Insbesondere werden sie eingesetzt zur Bildung der Zielvariablen, der Altersgruppen, den Franchisestufen und dem Indikator Spital-im-Vorjahr (siehe Tabelle 36). Zur Bildung der PCG werden detaillierte Medikamentenabrechnungen benötigt. Diese sind im Tarifpool der Sasis AG verfügbar.

Tabelle 36 Verfügbare Datensätze

	Gruppierungsebenen	Verfügbare Informationen	Verwendung
Datenpool Leistungsrecord	Alters- und Geschlechtsgruppe, Spital-im-Vorjahr (ab 2014), Versicherungsmodell, Franchisestufe, Unfalleinschluss (ja/nein), Schadenart, Leistungsart	Summe abgerechnete Leistungen (Kosten oder Anzahl Taxpunkte), Summe Konsultationen	Zähler der Zielvariablen; Erklärende Variablen Alter, Geschlecht, Spital-im-Vorjahr und Franchisestufe
Datenpool Erkranktenrecord	Alters- und Geschlechtsgruppe, Spital-im-Vorjahr (ab 2014)	Anzahl unterschiedliche Patienten pro Arzt und Jahr	Nenner der Zielvariablen
Tarifpool Leistungsrecord	Alters- und Geschlechtsgruppe, Tarifziffer: bei Medikamenten: Pharmacode	Anzahl der fakturierten Tarifpositionen	Bildung der pharmazeutischen Kostengruppen

Der Datenpool der Sasis AG ist aktuell die einzige Datenquelle, welche alle zu Lasten der OKP abgerechneten Leistungen erfasst. Ergänzt werden die Daten mit dem Tarifpool, welcher detaillierte Medikamentenabrechnungen enthält und so auch die Analysen von PCG zulässt.

Quelle: Eigene Darstellung.

In der Wirtschaftlichkeitsprüfung werden die Leistungsdaten dem Abrechnungsjahr zugeordnet. Diese Zuordnung hat den Vorteil, dass ein «Abrechnungsjahr» ein klar definiertes Beginn- und Enddatum hat. Eine Alternative wäre eine Zuordnung zum Behandlungsbeginnjahr. Da Patienten und Leistungserbringer jedoch das Recht haben, Rechnungen bis zu fünf Jahren nach der Behandlung bei der obligatorischen Krankenversicherung einzureichen, könnte ein Behandlungsbeginnjahr erst nach Ablauf dieser Frist wirklich abgeschlossen werden, was für die Wirtschaftlichkeitsprüfung ein erheblicher Nachteil ist. Jedoch ist in der Literatur nachgewiesen, dass Regressionsmodelle, bei welchen die erklärenden Variablen auf die frühere Leistungsinanspruchnahme abstützen (z. B. Vorjahreskosten, Spital-im-Vorjahr), bei einer Zuordnung zum Behandlungsbeginnjahr einen höheren Erklärungsgehalt haben als bei der Zuordnung zum Abrechnungsjahr (Beck 2013).

Eine Ausnahme bei der Zuordnung sind die Spitalaufenthalte zur Bildung der Variable «Spital-im-Vorjahr». Diese werden analog dem Risikoausgleich nach Behandlungsbeginnjahr gebildet. Für die vorliegende Analyse erwarten wir dadurch eher eine Verbesserung des Erklärungsgehalts, denn ambulante Leistungen, welche im Zusammenhang mit einem Spitalaufenthalt stattfanden, werden wohl teilweise im Folgejahr abgerechnet.

Im Rahmen des Projekts wurde getestet, ob eine zusätzliche Zuordnung der PCG zum Behandlungsbeginnjahr zu einer besseren Prognosegüte führen könnte. Diese Zuordnung der PCG wird ab 2020 auch im Risikoausgleich eingesetzt. In unseren Analysen wurde der Erklärungsgehalt bei der Zuordnung zum Behandlungsbeginnjahr jedoch verschlechtert. Dies ist deshalb nachvollziehbar, weil die auch Kosten (die Zielvariable) dem Abrechnungsjahr zugeordnet sind. Die Kosten der Medikamente, welche zur Zuordnung in die PCG führen, sind Teil der Zielvariablen, was den Erklärungsgehalt verbessert.

11.2.2 Massnahmen zur Sicherstellung der Anonymität der Leistungserbringer

Für die in den Projektzielen definierte praktische Umsetzung der Berechnungen war eine Übermittlung von Abrechnungsdaten an Polynomics notwendig. Die Daten mussten aber anonymisiert werden, damit Polynomics die Identität der Leistungserbringer nicht rekonstruieren konnte. Ein wichtiger Schritt dazu war die Anonymisierung der ZSR-Nummern. Zudem wurden die Charakteristika des Praxisstandortes nur soweit übermittelt, dass in jeder für Polynomics ersichtlichen geografischen Einheit mindestens fünf Ärzte pro Facharztgruppe vorhanden waren. Dies wurde einerseits durch die Übermittlung der Grossregion anstelle des Kantons erreicht und andererseits durch die Übermittlung einer limitierten Auswahl an Gemeindecharakteristika (Gemeinden gruppiert nach der Bevölkerungsdichte, und gruppiert nach der Sozialhilfequote).

11.2.3 Aggregation, Verknüpfungen und Ausschlüsse

Die Abrechnungsdaten zur Bildung der pharmazeutischen Kostengruppen (siehe nächster Abschnitt) stehen pro anonymisierten Veranlasser, Alter und Geschlecht zur Verfügung. Die anderen Datensätze werden ebenfalls auf dieser Stufe aggregiert, damit die Daten verknüpfbar sind. Die über 96-Jährigen wurden von ursprünglich fünf Altersgruppen zu einer Gruppe zusammengefasst.

Die Auswirkungen der Aggregationen sind in den ersten beiden Zeilen in Tabelle 37 dargestellt. Der Leistungsrecord wurde von ursprünglich über 35 Millionen Datensätzen auf rund 750‘000 Datensätze aggregiert. Diese stammen von 22‘273 anonymisierten ZSR-Nummern (Abrechnungsjahr 2015). Zusätzlich zum Leistungsrecord umfassten die Daten auch den «Erkranktenrecord». Dieser wird zur Bestimmung der Anzahl Erkrankter pro Arzt gebraucht (Nenner der Zielvariable) sowie zur Bildung des Morbiditätsindikators Spital-im-Vorjahr. Der Erkranktenrecord ist bereits auf der anonymen ZSR-Nummer, der AGG und Spital-im-Vorjahr aggregiert. Die Aggregation auf anonyme ZSR-Nummer und AGG reduzierte den Datensatz also nur um rund die Hälfte.

In den heutigen Verfahren werden Praxen nicht analysiert, welche unter 50 Erkrankte und pro Jahr unter CHF 100‘000 Bruttokosten haben (als Summe aus den erbrachten und veranlassten Kosten). Dadurch wurden weitere rund 4‘000 anonymisierte ZSR-Nummern ausgeschlossen. Um die Modelle berechnen zu können, mussten zudem Beobachtungen ausgeschlossen werden, bei welchen die Leistungen (Mittelwert pro AGG und Arzt) bei null lagen oder negativ waren. Zudem wurden Beobachtungen ausgeschlossen, bei welchen die Information zu der Grossregion nicht zur Verfügung standen.

Tabelle 37 Aggregationen, Verknüpfungen und Ausschlüsse

	Anzahl Beobachtungen Ursprungsdaten 2015	Anzahl Beobachtungen aggregierte Daten	Anzahl anonymisierte ZSR-Nummern
Leistungsrecord	35'321'750	752‘277	22‘273
Erkranktenrecord	1‘075‘516	606‘442	22‘395
Verknüpfte Datensätze		602‘885	22‘137
Ausschlüsse			
Ausschluss von Praxen mit weniger als 50 Erkrankten und unter CHF 100‘000 Bruttokosten		29‘252	4‘212
Ausschluss von Beobachtungen mit Kosten von null oder negativ		510	0
Ausschluss von Beobachtungen ohne Information zu Gemeindecharakteristika		1‘960	41
Bereinigter Datensatz		571‘163	17‘884

Nach der Aggregation, Verknüpfungen und Ausschlüssen stand ein Datensatz mit 17‘884 Praxen zur Verfügung.

Quelle: Daten der Sasis AG; Datenjahr 2015; eigene Berechnungen.

11.3 Regressionsdiagnostik

Im folgenden Abschnitt diskutieren wir ausgewählte Ergebnisse der Regressionsberechnung im Hinblick auf mögliche statistische Probleme.

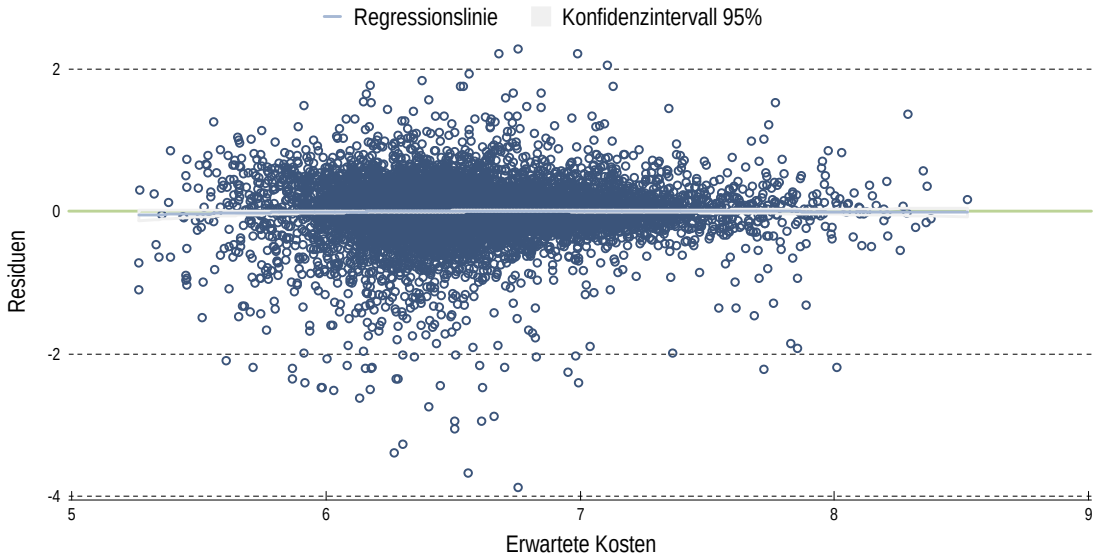
Ein wichtiger Aspekt der Regressionsdiagnostik ist die Verteilung der Residuen (ε_{it} in Gleichung 1, Abschnitt 4.1.1). Das Residuum ist definiert als die Abweichung der Vorhersage aus dem Modell vom tatsächlich beobachteten Wert. Ein positives Residuum bedeutet, dass das Modell die entsprechende Beobachtung unterschätzt. Beobachtungen mit negativen Residuen werden durch das Modell überschätzt.

11.3.1 Zusammenhang der Residuen und der erwarteten Werte

Besonders bedeutend ist die Analyse eines möglichen Zusammenhangs zwischen den aus dem Modell gemachten Vorhersagen und den Residuen. Existiert ein solcher Zusammenhang, könnte eine wichtige Modellannahme (Unabhängigkeit der Residuen von den erklärenden Variablen) verletzt sein.

In Abbildung 8 ist das Beispiel eines gut funktionierenden Modells dargestellt (Facharztgruppe: Kardiologen, logarithmiertes Modell). Auf der horizontalen Achse sind die Vorhersagen aus der Regression dargestellt. Links sind also Personengruppen, von denen geringe Kosten erwartet werden (z. B. junge Personen mit hohen Franchisen) während sich rechts Personen mit hohen erwarteten Kosten befinden. Auf der vertikalen Achse sind die Residuen dargestellt. Es ist kein Zusammenhang zwischen den beiden Variablen sichtbar.

Abbildung 8 Verteilung der Residuen, Zielvariable logarithmiert, Kardiologen



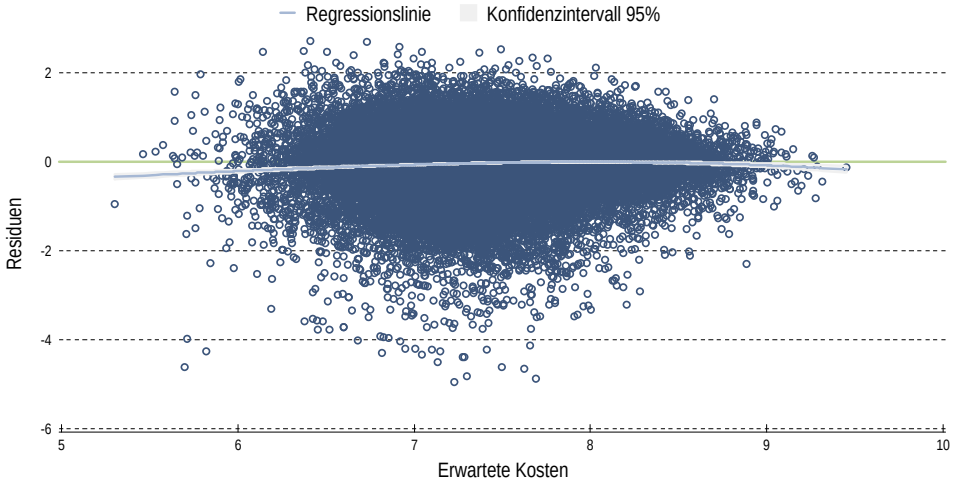
Kardiologen, jeder Punkt ist eine Beobachtung, N = 22'106, (unterschiedliche Ärzte = 369).

Auf der horizontalen Achse sind die Vorhersagen aus der Regression dargestellt. Links sind Personengruppen, von denen geringe Kosten erwartet werden (z. B. Junge mit hohen Franchisen) während sich rechts Personen mit hohen erwarteten Kosten befinden. Auf der vertikalen Achse sind die Residuen (Fehlerterme) dargestellt. Eine Regression hat gut funktioniert, wenn die Punkte zufällig verteilt sind. Im vorliegenden Fall der Kardiologen im logarithmierten Modell ist dies der Fall.

Quelle: Eigene Berechnungen, Polynomics.

Eine zweite Verteilung von Störtermen aus dem logarithmierten Modell ist in Abbildung 9 dargestellt. Gewählt wurde hier die Psychiatrie und Psychotherapie, weil bei dieser Facharztgruppe – besonders im nichttransformierten Fall – einige Probleme diskutiert werden können. Die Verteilung im transformierten Fall sieht ähnlich aus wie bei den Kardiologen. In beiden Fällen nehmen die Störterme im negativen Bereich eine deutlich grössere Spannweite ein als im positiven Bereich. Das Modell führt eher zu Überschätzungen als zu Unterschätzungen der Werte.

Abbildung 9 Verteilung der Residuen, Zielvariable logarithmiert, Psychiatrie und Psychotherapie



Auch in diesem Bild ist kaum ein Zusammenhang zwischen den Residuen und den erwarteten Werten erkennbar. Interessant ist jedoch, dass die Störterme im negativen Bereich eine grössere Ausdehnung haben als im positiven Bereich.

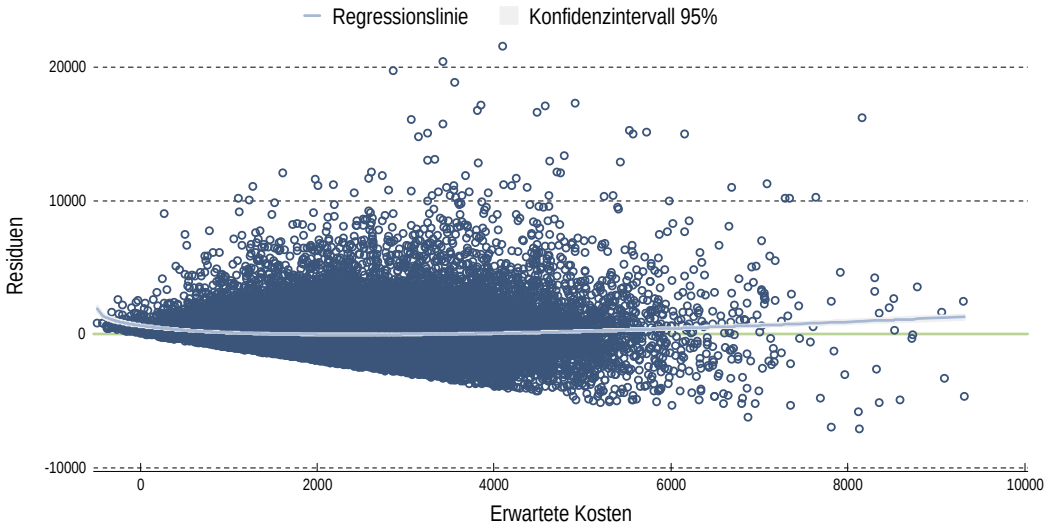
Quelle: Daten der Sasis AG; Datenjahr 2015; eigene Berechnungen.

In Abbildung 10 ist der Zusammenhang der geschätzten Werte mit den Residuen im untransformierten Modell dargestellt. Hier nehmen die positiven Störterme eine deutlich grössere Spannweite ein als die negativen. Dies ist nicht überraschend, denn die hohen Ausreisserwerte, welche in der deskriptiven Statistik beobachtet wurden, können durch das Modell nicht gut vorhergesagt werden.

Die Trendlinie (hellblauer Strich) liegt beim Hauptteil der Verteilung bei null, es ist generell also nicht von einem deutlichen Trend zu sprechen. An beiden Enden (Beobachtungen mit sehr geringen und mit sehr hohen erwarteten Kosten) werden die Kosten jedoch unterschätzt. Für eine Minderheit der Beobachtungen werden sogar negative Kosten vorhergesagt. Dieses Problem tritt häufig auf, wenn ein lineares Modell auf Daten angewendet wird, deren Verteilung an einer Seite «gestutzt» ist. Die «Stutzung» kommt daher, dass Kosten keine negativen Werte annehmen können, kleine Werte aber recht häufig auftreten. Dies führt zu einer «Verteilung mit einem abrupten Ende», wie sie auch in Abbildung 4 dargestellt ist. Lineare Modelle können in dieser Situation die kleinen Kosten nicht korrekt vorhersagen und «überschiessen» in den negativen Bereich. Die Stutzung ist auch der Grund dafür, dass im Bereich der negativen Residuen eine «Kante» sichtbar ist. Auf der Kante liegen Beobachtungen mit eher kleinen Werten, welche durch das Modell als zu hoch vorhergesagt werden. Wegen der Skala erscheinen ihre Residuen auf einer Linie, auch wenn sie sich unterscheiden.

Das Problem der Stutzung der Daten wurde in der gesundheitsökonomischen Literatur oft diskutiert (siehe unter anderem Duan et al., 1982; Buntin und Zaslavsky, 2004; Beck, 2013). Trotz dieser Probleme werden untransformierte lineare Modelle in der Praxis häufig angewendet. Die praktische Erfahrung zeigt, dass sie oft eine deutlich bessere Prognosegüte aufweisen als andere Schätzmethoden wie beispielsweise Two-Part-Modelle oder GLM (Buntin und Zaslavsky, 2004; Beck, 2013).

Abbildung 10 Verteilung der Residuen, untransformiert



Psychiatrie und Psychotherapie, N = 52‘826, Ärzte = 2'125.

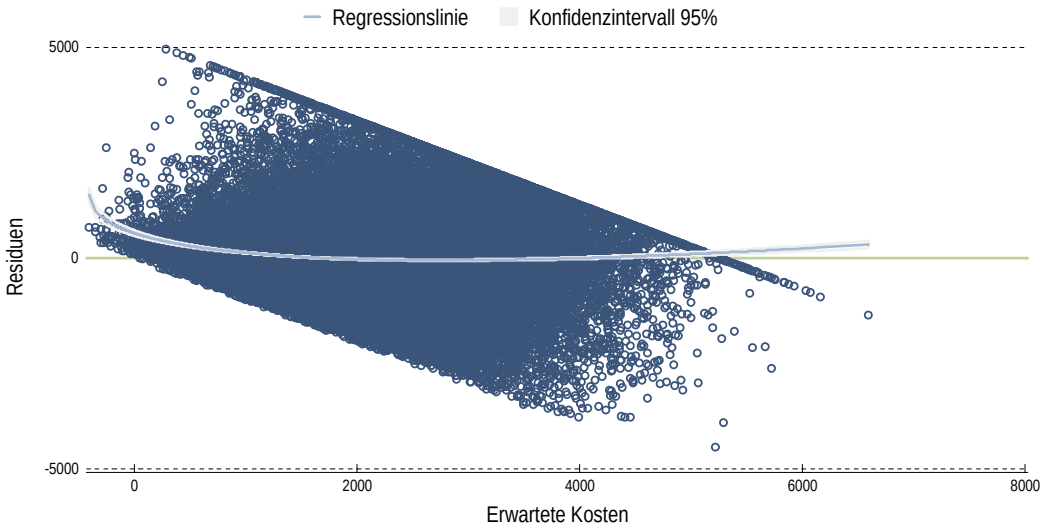
In einem untransformierten Modell gibt es eine Minderheit an Beobachtungen, für die das Modell die echten Werte stark unterschätzt. Diese haben hohe positive Residuen. Bei den negativen Residuen ist eine Kante sichtbar. Diese entsteht, weil es in den Daten viele Beobachtungen mit kleinen Werten gibt. Die kleinen Werte werden durch das Modell überschätzt.

Quelle: Daten der Sasis AG; Datenjahr 2015; eigene Berechnungen.

Wie bereits erwähnt gibt es in einem untransformierten Modell wie in Abbildung 10 eine kleine Gruppe von hohen Ausreissern, welche nicht gut erklärt wird und daher hohe Residuen hat. Werden die Originaldaten winsorisiert, kann der Einfluss dieser Gruppe reduziert werden. In Abbildung 11 ist der Einfluss der Winsorisierung auf die Residuen dargestellt. Er wirkt ähnlich wie die «Stutzung» am unteren Ende, aber auf die hohen Residuen. Die sind nun durch eine deutliche Schranke begrenzt.

Bei einem Vergleich von Abbildung 10 und Abbildung 11 ist zu beachten, dass sich die Skalen verändert haben. Die Y-Achse in Abbildung 11 hat weniger stark positive Werte, was ein klarer Effekt der Winsorisierung ist. Bemerkenswert ist, dass auch die X-Achse nicht die gleichen Werte umfasst. Durch die Stutzung der Ausreisser in der Schätzung nehmen die Vorhersagen eine kleinere Spannbreite ein. Dies weist darauf hin, dass einige Koeffizienten, insbesondere einige Praxiseffekte, im unwinsorisierten Modell deutlich grösser sind als im winsorisierten. Die fünf Prozent höchsten Werte üben also einen Einfluss auf den spezifischen Praxiseffekt aus.

Abbildung 11 Verteilung der Residuen, Zielvariable winsorisiert, Psychiatrie und Psychotherapie



Psychiatrie und Psychotherapie, N = 52‘826, Ärzte = 2'125.

Der Einfluss der hohen Residuen kann durch die Winsorisierung reduziert werden. Die hohen Residuen werden gestutzt. Wie beim Vergleich der X-Achse mit Abbildung 10 deutlich wird, gibt es auch Beobachtungen, für die dadurch geringere erwartete Werte berechnet werden.

Quelle: Daten der Sasis AG; Datenjahr 2015; eigene Berechnungen.

11.3.2 Heteroskedastie

Heteroskedastie führt nicht zu einer Verzerrung der Schätzer, diese können jedoch unpräzise sein. Wie stark das Problem der Heteroskedastie ist, kann beispielsweise mittels eines Breusch-Pagan-Tests überprüft werden (siehe z. B. Wooldridge, 2016). Dazu wird die Varianz der Residuen auf die erklärenden Variablen regressiert. Mit dem einen F-Test kann dann geprüft werden, ob die Variablen gemeinsam signifikant sind.

Im Folgenden führen wir den Breusch-Pagan-Test für zwei unterschiedliche Modelle durch. Erstens für ein Modell mit allen erklärenden Variablen und den praxisspezifischen Effekten sowie zweitens für ein Modell nur mit dem praxisspezifischen Effekt. Die Resultate sind in Tabelle 38 dargestellt. Das Modell 1, in welchem die Varianz mit allen Variablen berechnet wird, erreicht ein beachtliches R^2 von 10 bis 20 Prozent. Der F-Test gibt an, dass die erklärenden Variablen gemeinsam statistisch signifikant sind. Es liegt klar Heteroskedastie vor.

Im zweiten Modell, wo nur noch die praxisspezifischen Effekte enthalten sind, ist der Erklärungsgehalt wesentlich geringer. Der Zusammenhang zwischen der Varianz der Residuen und den praxisspezifischen Effekten ist nicht sehr stark, aber trotzdem vorhanden. Ausser bei den Ärzten der Kinder- und Jugendmedizin gibt der F-Test an, dass die praxisspezifischen Effekte gemeinsam signifikant zur Erklärung der Fehler beitragen. Das Problem der Heteroskedastie ist auch in Bezug auf die Praxiseffekte vorhanden.

Tabelle 38 Breusch-Pagan-Test für Heteroskedastie in den Störtermen

	Allgemeine Innere	Chirurgie	Gynäkologie	Kardiologie	Kinder/ Jugend	Ophthal-mologie	Psychiatrie/ Psycho-therapie
N	196'637	9'687	22'106	12'780	16'135	30'873	52'826
N Ärzte	5'154	278	1'137	369	941	770	2'125
Adj. R^2 Modell 1	0.11	0.13	0.18	0.10	0.14	0.21	0.10
F-Test Modell 1	5.5***	5.2***	5.0***	4.3***	3.6***	10.9***	3.7***
Adj. R^2 Modell 2	0.05	0.06	0.004	0.02	-0.02	0.17	0.02
F-Test Modell 2	3.1***	3.0***	1.1**	1.8***	0.6	9.0***	1.4***

Die ausgewiesenen R^2 beziehen sich auf die Hilfsregression zur Berechnung des Breusch-Pagan-Tests.

Modell 1: alle erklärenden Variablen inkl. praxisspezifisch Effekte; Modell 2: nur praxisspezifische Effekte.
Der F-Test testet die gemeinsame Signifikanz der praxisspezifischen Effekte.
Signifikanzniveaus: *** p<0.01, ** p<0.05, * p<0.1.

Die Zielvariablen in dieser Regression ist die Residualvarianz. So kann getestet werden, ob die erklärenden Variablen (Modell 1) oder die spezifischen Praxiseffekte (Modell 2) mit der Varianz der Residuen korreliert sind (Heteroskedastie). In beiden Fällen ist Heteroskedastie vorhanden, bei den Praxiseffekten ist sie jedoch eher schwach (mit Ausnahme der Ophthalmologen).

Quelle: Daten der Sasis AG; Datenjahr 2015; eigene Berechnungen.

Auffallend ist, dass bei den Ärzten der Ophthalmologie sowohl das R^2 als auch der F-Test wesentlich höhere Werte annehmen als bei den anderen Ärzten. In dieser Facharztgruppe ist die Residualvarianz bei gewissen Ärzten viel höher als bei anderen. Dies deutet darauf hin, dass es kein zufälliger Fehler ist, sondern der Grund in Praxischarakteristika liegt, welche wir nicht beobachten. Dieses geht darum in den praxisspezifischen Effekt ein. Die Ophthalmologie ist daher eine Facharztgruppe, bei welcher im dritten Teilprojekt abgeklärt werden müsste, ob spezifische erklärende Variablen (z. B. Tarmedpositionen) gefunden werden können.

11.4 Einbezug des Praxisstandortes

Wie in Abschnitt 7.2.2 beschrieben haben wir unterschiedliche Spezifikationen der Gemeindecharakteristika getestet. Die entsprechenden Koeffizienten sind in Tabelle 39 dargestellt.

Tabelle 39 Koeffizienten der Praxischarakteristika

Charakteristika	Koeffizient	Charakteristika	Koeffizient	Charakteristika	Koeffizient
Sozialhilfequote Gruppe 2	-0.02 (-0.84)	Einwohner-dichte Gruppe 2	-0.02 (-0.74)	Ausländer-anteil Gruppe 2	-0.05 (-0.89)
Sozialhilfequote Gruppe 3	-0.01 (-0.39)	Einwohner-dichte Gruppe 3	-0.02 (-0.97)	Ausländer-anteil Gruppe 3	-0.03 (-0.69)
Sozialhilfequote Gruppe 4	-0.01 (-0.69)	Einwohner-dichte Gruppe 4	-0.02 (-0.86)	Ausländer-anteil Gruppe 4	-0.00 (-0.05)
Sozialhilfequote Gruppe 5	-0.02 (-0.92)	Einwohner-dichte Gruppe 5	0.02 (0.89)	Ausländer-anteil Gruppe 5	0.01 (0.13)
Sozialhilfequote Gruppe 6	0.04 (1.59)				

Modell mit Grossregionen, Standardfehler in Klammern.

Die getesteten Indikatoren Sozialhilfequote, Einwohnerdichte und Ausländeranteilsgruppe hatten keinen signifikanten Einfluss auf den Praxiseffekt. Die Referenzgruppe ist die Gruppe mit dem geringsten Wert. Die Gruppe mit dem höchsten Wert hat jeweils einen positiven Koeffizienten. Eine hohe Sozialhilfequote in der Gemeinde führt bspw. zu einem rund 4% höheren Indexwert.

Quelle: Daten der Sasis AG; Datenjahr 2015; eigene Berechnungen.

11.5 Datenaufbereitungen der Versichererdaten

Die Individualdaten wurden aus den vier Datensätzen direkte Kosten, veranlasste Kosten, Patientenrecord und dem Medikamentenrecord erstellt. Diese wurden jeweils einzeln aufbereitet und anschliessend zusammengefügt. Die Daten lagen für die Jahre 2013, 2014 und 2015 vor.

Tabelle 40 Teildatensätze der Versichererdaten und enthaltene Variablen

Direkte Kosten	Veranlasste Kosten	Patientenrecord	Medikamentenrecord
▪ Abrechnungsjahr ▪ Kanton ▪ Leistungsart ▪ Partnerart_UG ▪ PatientenID ▪ RechnungsstellerID ▪ Fakturabetrag	▪ Abrechnungsjahr ▪ Leistungsart ▪ Fakturabetrag ▪ Partnerart_OG ▪ PatientenID ▪ RechnungsstellerID ▪ VeranlasserID	▪ Geburtsjahr ▪ Geschlecht ▪ Franchise ▪ Versicherungsmodell ▪ Spitalaufenthalt ▪ Behandlungsbeginnjahr ▪ PatientenID	▪ Behandlungsjahr ▪ PatientenID ▪ RechnungsstellerID ▪ PCG ▪ Anzahl DDD ▪ VeranlasserID

Quelle: Abrechnungsdaten von drei Krankenversicherern; eigene Darstellung.

Bei den direkten Kosten wurden alle negativen Kosten (Fakturabetrag) mit null ersetzt (1'174 Fälle). Zudem wurden die Kosten für die Arztbehandlung mit dem durchschnittlichen Taxpunktwert standardisiert. Pro Rechnungssteller konnte nur eine Partnerart (Facharztgruppe) berücksichtigt werden. Bei Rechnungssteller, die für den gleichen Patienten und die gleiche Leistungsart im gleichen Jahr verschiedenen Partnerarten aufweisen, wurden der gesamte Rechnungsbetrag des Patienten der Partnerart mit dem grösseren Fakturabetrag zugewiesen (102 Fälle). Wenn der Fakturabetrag genau gleich hoch war, wurde die Allgemeine Innere Medizin als Facharztgruppe verwendet (2 Fälle). Bei unterschiedlichen Leistungsarten wurde ebenfalls die Facharztgruppe Allgemeine Innere Medizin verwendet (41 Fälle). Wenn ein Rechnungssteller für unterschiedliche Patienten unterschiedliche Leistungsarten aufwies, wurde die Facharztgruppe mit dem grösseren Anteil der Patienten gewählt (522 Fälle).

Bei den veranlassten Kosten wurden ebenfalls alle negativen Kosten (Fakturabetrag) mit null ersetzt (1'147 Fälle). Anschliessend wurde die veranlassten Kosten mit den direkten Kosten zusammengefügt und die veranlassten Kosten zu den Gesamtkosten pro Arzt und Patient addiert. Für das Zusammenfügen der beiden Datensätze wurden die folgenden Variablen verwendet:

- Direkte Kosten: Abrechnungsjahr, RechnungsstellerID, PatientenID
- Veranlasste Kosten: Abrechnungsjahr, VeranlasserID, PatientenID

Tabelle 41 Zusammenfügen der direkten und veranlassten Kosten

Beobachtungen	2013	2014	2015
nur direkte Kosten	273'804	275'499	280'977
direkte und veranlasste Kosten	217'878	232'144	237'507
nur veranlasste Kosten	64'602	65'168	61'821
total verwendet	491'682	507'643	518'484

Quelle: Abrechnungsdaten von drei Krankenversicherern; eigene Berechnungen.

Im dritten Schritt wurden pro Patient die Daten aus dem Patientenrecord zu den Kostendaten hinzugefügt. Dazu wurden die folgenden Variablen verwendet:

- Kosten: Abrechnungsjahr, PatientenID_anonym
- Patientenrecord: Behandlungsbeginnjahr, PatientenID_anonym

Wenn keine Patientendaten vorlagen, wurden die Patientendaten aus dem Vorjahr verwendet.

Tabelle 42 Zusammenfügen der Kosten- und Patientendaten

Beobachtungen	2013	2014	2015
nur Kosten	5'973	9'557	14'211
Kosten und Patientenrecord	485'709	498'086	504'273
Kosten und Patientenrecord Vorjahr	4'818	7'997	12'540
nur Patientenrecord	173'830	170'113	145'751
total verwendet	490'527	506'083	516'813

Quelle: Abrechnungsdaten von drei Krankenversicherern; eigene Berechnungen.

Im letzten Schritt wurden noch der Medikamentenrecord zu den Daten hinzugefügt. Dazu wurden die folgenden Variablen verwendet:

- Kosten: Abrechnungsjahr, RechnungsstellerID, PatientenID
- Medikamentenrecord: Behandlungsbeginnjahr, VeranlasserID, PatientenID

Insgesamt stehen damit pro Jahr rund 500'000 Beobachtungen zur Verfügung.

Tabelle 43 Zusammenfügen der Kostendaten mit dem Medikamentenrecord

Beobachtungen pro Patient und Arzt	2013	2014	2015
ohne PCG	365’420	375’777	388’951
mit PCG	125’107	130’306	127’862
total verwendet	490'527	506'083	516'813

Quelle: Abrechnungsdaten von drei Krankenversicherern; eigene Berechnungen.

12 Quellenverzeichnis

Adams, J.L. 2009. The Reliability of Provider Profiling. Product Page. Santa Monica: RAND Corporation.

Adams, J.L., A. Mehrotra und E.A. McGlynn. 2010. Estimating Reliability and Misclassification in Physician Profiling. Santa Monica: RAND Corporation.

Adams, J.L., A. Mehrotra, J.W. Thomas und E.A. McGlynn. 2010. Physician Cost Profiling – Reliability and Risk of Misclassification. Detailed Methodology and Sensitivity Analyses (Technical Appendix). Santa Monica: RAND Corporation.

Beck, K. 2013. *Risiko Krankenversicherung – Risikomanagement in einem regulierten Krankenversicherungsmarkt*. 3. Auflage. Bern: Haupt Verlag.

Beck, N. 2011. Of Fixed-Effects and Time-Invariant Variables. *Political Analysis*, 19(02):119–122. doi:10.1093/pan/mpr010.

Buntin, M.B. und A.M. Zaslavsky. 2004. Too much ado about two-part models and transformation? *Journal of Health Economics*, 23(3):525–542. doi:10.1016/j.jhealeco.2003.10.005.

Cameron, A.C. und D.L. Miller. 2015. A practitioner's guide to cluster-robust inference. *Journal of Human Resources*, 50(2):317–372.

Duan, N., W.G. Manning, C. Morris und J.P. Newhous. 1982. A comparision of Alternative Models for the Demand for Medical Care. Santa Monica, CA: Rand Cooperation.

Eijkenaar, F. und R.C.J.A. van Vliet. 2013. Profiling Individual Physicians Using Administrative Data From a Single Insurer: Variance Components, Reliability, and Implications for Performance Improvement Efforts. *Medical Care*, 51(8):731–739. doi:10.1097/MLR.0b013e3182992bc1.

Eijkenaar, F. und R.C.J.A. van Vliet. 2014. Performance Profiling in Primary Care: Does the Choice of Statistical Model Matter? *Medical Decision Making*, 34(2):192–205. doi:10.1177/0272989X13498825.

Gardiol, L., P.-Y. Geoffard und C. Grandchamp. 2005. Separating Selection an Incentive Effects in Health Insurance. CEPR Discussion Paper (5380).

Kaiser, B. 2016. Methodische Weiterentwicklung der Wirtschaftlichkeitsprüfung. Basel: B,S,S. Volkswirtschaftliche Beratung AG.

Kauer, L. 2016. Long-term Effects of Managed Care: Long-term Effects of Managed Care. *Health Economics*. doi:10.1002/hec.3392.

Kristensen, T., K.R. Olsen, H. Schroll, J.L. Thomsen und A. Halling. 2014. Association between fee-for-service expenditures and morbidity burden in primary care. *The European Journal of Health Economics*, 15(6):599–610. doi:10.1007/s10198-013-0499-7.

Manning, W.G. 1998. The Logged Dependent Variable, Heteroscedasticity, and the Retransformation Problem. *Journal of Health Economics*, 17(3):283–95.

Manning, W.G. und J. Mullahy. 2001. Estimating log Models: To Transform or Not to Transform? *Journal of Health Economics*, 20(4):461–94.

Mihaylova, B., A. Briggs, A. O'Hagan und S.G. Thompson. 2011. Review of Statistical Methods for Analysing Healthcare Resources and Costs. *Health Economics*, 20(8):897–916. doi:10.1002/hec.1653.

Plümper, T. und V.E. Troeger. 2007. Efficient Estimation of Time-Invariant and Rarely Changing Variables in Finite Sample Panel Analyses with Unit Fixed Effects. *Political Analysis*, 15(02):124–139. doi:10.1093/pan/mpm002.

Pope, G.C., R.P. Ellis, A.S. Ash, C.-F. Liu, J.Z. Ayanian, D.W. Bates, H. Burstin, L.I. Iezzoni und M.J. Ingber. 2000. Principal Inpatient Diagnostic Cost Group Model for Medicare Risk Adjustment. *Health Care Financing Review*, 21(3):93–118.

Roth, H.-R. und W. Stahel. 2005. Die ANOVA-Methode zur Prüfung der Wirtschaftlichkeit von Leistungserbringern nach Artikel 56 KVG. Zürich: Seminar für Statistik, ETH Zürich.

von Rotz, S., U. Kunze und K. Beck. 2008. Der Ärzteindex - Ein instrument zur Beurteilung der Wirtschaftlichkeit von Grundversorgern. *Gesundheitsökonomie und Qualitätsmanagement*, 13:142–148. doi:10.1055/s-2007-963626.

Schira, J. 2009. *Statistische Methoden der VWL und BWL: Theorie und Praxis*. Pearson Deutschland GmbH.

Schmid, C. und K. Beck. 2015. Wirken hohe Franchisen kostendämpfend? *Schweizerische Ärztezeitung*, 96(35):1238–1239.

Schmidt, P. und R.C. Sickles. 1984. Production Frontiers and Panel Data. *Journal of Business and Economic Statistics*:367–374.

Schwenkglenks, M. 2010. Vergleich verschiedener Instrumente (Rechnungsstellerstatistik der santésuisse und Praxisspiegelder Trustcenter) zur Beurteilung der von Schweizer Ärzten in der Grundversorgung verursachten Behandlungskosten. Basel: Institute of Pharmaceutical Medicine / ECPM, Universität Basel.

Thomas, J.W., K. Grazier und K. Ward. 2004a. Comparing Accuracy of Risk-Adjustment Methodologies Used in Economic Profiling of Physicians. *Inquiry*: 218–231.

---. 2004b. Economic Profiling of Primary Care Physicians: Consistency among Risk-Adjusted Measures. *Health Services Research*, 39(4):985–1004.

Thomas, J.W. und K. Ward. 2006. Economic profiling of physician specialists: use of outlier treatment and episode attribution rules. *INQUIRY: The Journal of Health Care Organization, Provision, and Financing*, 43(3):271–282.

Trottmann, M., H. Telser, D. Stämpfli, P.K.E. Hersberger, K. Matter und M. Schwenkglenks. 2015. Übertragung der niederländischen PCG auf Schweizer Verhältnisse. Olten: Bundesamt für Gesundheit BAG.

Van de Ven, W. und R.P. Ellis. 2000. Risk adjustment in competitive health plan markets. In: *Handbook of Health Economics*, 14:755–845. Volume 1 A. Elsevier.

Wasem, J., G. Lux und H. Dahl. 2010. Beurteilung des Screening - Verfahrens der Krankenversicherer in der Schweiz zur Identifikation von Überarztung. Gutachten beauftragt von: Verein Ethik und Medizin Schweiz. Essen: ForBig - Forschungsnahe Beratungsgesellschaft im Gesundheitswesen GmbH.

White, H. 1980. A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity. *Econometrica*, 48(4):817–838. doi:10.2307/1912934.

Wooldridge, J.M. 2010. *Econometric Analysis of Cross Section and Panel Data*. Bd. 1.

Wooldridge, J.M. 2016. *Introductory econometrics: a modern approach*. 6. ed. Boston, Mass.: Cengage Learning.

Polynomics AG
Baslerstrasse 44
CH-4600 Olten

www.polynomics.ch
polynomics@polynomics.ch

Telefon +41 62 205 15 70
Fax +41 62 205 15 80