Statistische Bewertung unterschiedlicher Hierarchievarianten im Klassifikationssystem für den Risikostrukturausgleich

Gutachten im Auftrag des Bundesversicherungsamtes

Prof. Dr. Thomas Schäfer

Oberuhldingen, im April 2011

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Zusammenfassung

(1) Im Klassifikationssystem des morbiditätsorientierten Risikostrukturausgleichs werden die Versicherten Morbiditätsgruppen zugeordnet, die in Hierarchien zusammengefasst sind. Dabei bestehen für einige Krankheitshierarchien unterschiedliche Möglichkeiten der Ausgestaltung, welche aus Sicht der medizinischen Klassifikation den gleichen Grad der Plausibilität aufweisen. Üblicherweise wird dann diejenige Ausgestaltungsvariante ausgewählt, welche bei der regressionsanalytischen Berechnung der Zuschläge mit einem höheren Anteil der erklärten Varianz (R^2) verbunden ist. Da von der Umgestaltung einer Hierarchie in der Regel jedoch nur vergleichsweise wenige Versicherte der Versichertenstichprobe betroffen sind, unterscheiden sich die zugehörigen R^2-Werte, die von der großen Zahl der „Nichtbetroffenen“ dominiert werden, vielfach erst in den hinteren Nachkommastellen. Die Entscheidung für die eine oder andere Variante auf der Basis der Differenz der R^2-Werte erscheint daher als wenig belastbar. Aufgabe des Gutachters war vor diesem Hintergrund, ein quantitatives Bewertungsverfahren zu entwickeln, das zu belastbareren Entscheidungen führt. Das im Folgenden vorgeschlagene Bewertungsverfahren ist am Beispiel von je drei Ausgestaltungsvarianten der Hierarchien 02 (Neubildungen), 04 (Metabolische Erkrankungen) und 19 (Erkrankungen der Lunge) erprobt und optimiert worden. Die hierfür erforderlichen Daten wurden einer Sonderauswertung des Bundesversicherungsamtes für das Gutachten entnommen.

(2) Ein möglicher Ansatz das bisher praktizierte Verfahren zu verbessern, besteht darin, andere Maßzahlen zu verwenden als das Bestimmtheitsmaß R^2, das bekanntlich äußerst empfindlich auf statistische Ausreißer – d. h. im Kontext auf besonders teuere Versicherte – reagiert. Der naheliegende Gedanke, Vorhersageverhältnisse (Predictive Ratios – PR) für die den hierarchischen Morbiditätsgruppen (HMG) der Hierarchie zugeordneten Versichertengruppen heranzuziehen, lässt sich aus innermathematischen Gründen nicht verwirklichen, weil die PR von Versichertengruppen, deren Indikatorvariablen als Prädiktoren im Regressionsmodell Verwendung finden, stets den Wert 1 hat.

Im Rahmen eines von Cumming et al. (2002) durchgeführten empirischen Vergleichs der zum Zwecke der Risikoadjustierung in den USA eingesetzten Versichertenklassifikationssysteme wurde neben R^2 auch der mittlere absolute Vorhersagefehler (Mean Absolute Prediction Error – MAPE) und eine darauf aufbauende, von Cumming vorgeschlagene Maßzahl berechnet, die sich von R^2 nur dadurch unterscheidet, dass anstelle der Quadrate die Absolutbeträge verwendet werden. Diese von den Autoren der Vergleichsstudie als „Cumming’ s Prediction Measure“ (CPM) bezeichnete Maßzahl vermeidet die o. g. Nachteile von R^2. Cumming bevorzugte CPM als relative Maßzahl gegenüber der auf einer absoluten Skala in Dollar bzw. Euro messenden MAPE, weil sich die von ihm untersuchten Klassifikationssysteme erheblich in der Gesamtvarianz bzw. Totalvariation der zugehörigen Regressionsmodelle unterscheiden. Dies ist aber bei der hier diskutierten Fragestellung gänzlich anders, wie im Gutachten im Einzelnen dargelegt wird. Zwei Ausgestaltungsvarianten einer Hierarchie führen stets auf den gleichen Nenner von

CPM, so dass Entscheidungen auf der Basis von MAPE gleichlaufend mit denjenigen auf der Basis von CPM ausfallen.

Im empirischen Teil des vorliegenden Gutachtens hat sich gezeigt, dass die Verwendung des mittleren absoluten Vorhersagefehlers erhebliche Interpretationspotenziale eröffnet. In zwei der sechs angestellten Vergleiche kommt man auf der Basis der Differenz der MAPE-Werte zum entgegengesetzten Ergebnis wie auf der Basis der Differenz der R^2-Werte. Im Rahmen der vergleichenden Bewertung von Ausgestaltungsvarianten einer Hierarchie wird daher die Verwendung von MAPE als weiteres Maß für die Güte der Anpassung neben R^2 empfohlen.

Will man über die jährliche Betrachtung hinaus auch die Konsistenz der Entscheidungen bei der Ausgestaltung einer Hierarchie über mehrere Jahre hinweg prüfen, so ist zu beachten, dass sich die Nenner in verschiedenen Jahren unterscheiden. Für solche Zwecke sollte die absolute Maßzahl MAPE dann zur Bewertung ihre Größe durch die relative Maßzahl CPM flankiert werden.

(3) Um dem oben erwähnten Verdünnungseffekt infolge einer großen Zahl von Versicherten, die von den Unterschieden in der Ausgestaltung einer Hierarchie nicht betroffen sind, entgegen zu wirken, wird vorgeschlagen – bezogen auf eine bestimmte vorgegebene Hierarchie mit zwei verschiedenen Varianten – aus der Menge der nichtbetroffenen Versicherten eine Unterstichprobe (ohne Zurücklegen) zu ziehen und zusammen mit der Gesamtheit der Versicherten der Betroffenen auszuwerten, so dass das Design in Richtung eines balancierten Designs verändert wird. Der Stichprobenumfang der Nichtbetroffenen (also der Unterstichprobe) sollte der Zahl der Betroffenen entsprechen. Allerdings sollte er nicht unter 2.000 fallen und ggf. entsprechend gesetzt werden. Die Unterstichprobe, zusammen mit den betroffenen Versicherten, bildet dann einen Datensatz, der als „Bewertungsstichprobe“ bezeichnet wird. Werden zwei Varianten der Ausgestaltung einer Hierarchie miteinander verglichen (wobei die erste die herkömmliche „Basisvariante“ darstellen möge), so lassen sich für jede der beiden Varianten auch ohne Neukalibrierung des Modells aus der Bewertungsstichprobe verschiedene Maßzahlen berechnen. Die Regressionskoeffizienten stammen dabei aus der Kalibrierung des Modells in der jeweiligen Variante an der vollen Versichertenstichprobe. Definiert man

$$(*) \qquad D_1 = R_2^2 - R_1^2, \quad D_2 = CPM_2 - CPM_1, \; D_3 = MAPE_2 - MAPE_1, \quad D_4 = r_2 - r_1,$$

wobei r die Korrelation zwischen den tatsächlichen und den vorhergesagten Ausgaben darstellt, so wird vorgeschlagen, dass sich die Bewertung auf die Verteilungen dieser vier Differenzen stützt. Die Korrelation r wurde als vierte Maßzahl hinzu genommen, weil sie einem vertrauten statistischen Konzept folgt, gut interpretierbar ist und weil ihr Quadrat nicht mit R^2 übereinstimmt, wenn beide Maßzahlen aus einer Bewertungsstichprobe berechnet werden. Die Differenz der CMP-Werte wird nur für eine Betrachtung im Längsschnitt über mehrere Jahre benötigt und kann bei Beschränkung auf ein Jahr weggelassen werden.

Wenn sich die zu vergleichenden Varianten in der Zahl der Prädiktoren unterscheiden, so wird zweckmäßiger Weise noch eine fünfte Differenz hinzugezogen,

$$(**) \quad D_5 = BIC_2 - BIC_1,$$

die auf dem Bayesschen Informationskriterium (BIC) basiert, welches einen Strafterm für zusätzliche Prädiktoren enthält (zu Details s. Abschnitt 2.4.3).

Da die Verteilungen der genannten Differenzen nicht bekannt sind, müssen sie geschätzt werden, wobei ein Ansatz vorgeschlagen wird, der dem Begriff „Resampling“ subsumiert werden kann.

Hierzu ist die Ziehung der beschriebenen Unterstichprobe aus den Nichtbetroffenen, die ja ihrerseits ohne Zurücklegen gezogen wurde, nun n-mal mit Zurücklegen zu wiederholen (d. h. Versicherte, die in vorangegangen Ziehungen in die Unterstichprobe gelangt sind, können wieder gezogen werden). Auf diese Weise entstehen n Bewertungsstichproben mit jeweils n Ausprägungen für die drei zur Bewertung herangezogenen Differenzen. Deren empirischen Verteilungen lassen sich dann z. B. durch Darstellung von Histogrammen oder Berechnung von Mittelwerten und Perzentilen auswerten.

(4) Bei der Erprobung des Verfahrens wurden zum Zwecke der Optimierung zwei Variationsmöglichkeiten vorgesehen. Eine der beiden betraf die Definition der Betroffenen. Es gab eine enge und eine weite Operationalisierung dieses Begriffs. Bei Verwendung der letzteren wurden Versicherte dann als Betroffene eingestuft, wenn sie nur irgendeiner der hierarchischen Morbiditätsgruppen zugeordnet waren, die in den beiden verglichenen Ausgestaltungsvarianten zusammen vorkamen. Bei der engen Operationalisierung wurden nur diejenigen Versicherten als Betroffene eingestuft, die hierarchischen Morbiditätsgruppen zugeordnet waren, deren Definition oder Einfügung in die Hierarchie sich zwischen den zu vergleichenden Varianten unterschieden.

Die zweite Variationsmöglichkeit betraf die Ausbalancierung des Designs. Der Stichprobenumfang der Unterstichprobe aus den Nichtbetroffenen konnte als q-faches der Zahl der Betroffenen gewählt werden (q =1, 2, 3).

Als Ergebnis der Erprobung lässt sich festhalten, dass die enge Operationalisierung der „Betroffenen“ und die Wahl von q=1 dem Zweck des Designs, der Verdünnung der Effekte entgegenzuwirken, am besten entspricht. Aber da die Differenz der mittleren absoluten Vorhersagefehler in zwei der sechs Vergleiche beim Übergang von der weiten zu der engen Operationalisierung der „Betroffenen“ das Vorzeichen gewechselt hat, sollte das Verfahren stets mit beiden Operationalisierungen durchgeführt werden, um für die Bewertung über alle relevanten Informationen verfügen zu können.

(5) Das Verfahren hat sich als außerordentlich stabil erwiesen. Für die Erprobung wurden jeweils insgesamt n=9.000 Bewertungsstichproben ausgewertet. Das Vorzeichen der Differenzen D_1, D_2 und D_3 aus (*) hat in keinem der betrachteten Fälle innerhalb dieser 9.000 Wiederholungen einen Wechsel erfahren. Es hat sich darüber hinaus gezeigt, dass die Mittelwerte der Verteilungen der Differenzen selbst bei Zugrundelegung der

engen Operationalisierung von „betroffen“ schon nach einer kleinen Zahl von Wiederholungen bei Erhöhung des n allenfalls noch in der zweiten signifikanten (d. h. von Null verschiedenen) Nachkommastelle geringfügige Schwankungen aufweisen. Wenn in den zukünftigen Anwendungen etwa n = 100 gewählt wird, bleiben für eine stabile Berechnung der empirischen Verteilungen der Differenzen ausreichend Reserven.

(6) Betrachtet man den Effekt des Verfahrens auf die Differenzen der R^2-Werte, so ist das Vorzeichen in allen Bewertungsstichproben das gleiche, das die jeweilige Differenz bei Berechung aus der vollen Versichertenstichprobe aufweist. Allerdings sind die Mittelwerte aus den Bewertungsstichproben um einen Faktor größer, der sich für n=9.000 bei enger Operationalisierung von „betroffen“ und Wahl von q=1 in der Erprobung je nach Art der zu bewertenden Hierarchievarianten zwischen und 22,6 und 2.226,2 bewegt. Für die CPM-Differenzen (eher mit den R^2-Differenzen vergleichbar als die MAPE-Differenzen) sind die Faktoren noch größer und variieren im absoluten Betrag zwischen 47,5 und 6.596,7. Allerdings trat bei zwei der Vergleiche auf der Basis von CPM bzw. MAPE – wie bereits berichtet – ein interpretationsbedürftiger Vorzeichenwechsel gegenüber dem Wert der Differenz bei Berechnung aus der vollen Stichprobe auf.

(7) Auch wenn der mittlere absolute Fehler im Vergleich zu R^2 große Vorzüge aufweist, erscheint es nicht als zweckmäßig, auf die Berechnung von R^2 zu verzichten. Die Stützung der Bewertungsentscheidung auf mehrere Maßzahlen und zwei verschiedene Betroffenkonzepte eröffnet allein durch das Studium der übereinstimmenden bzw. abweichenden Bewertungen erhebliche Interpretationspotenziale.

Schließlich sehen die Krankenkassen der Frage, welche Ausgestaltungsvariante einer Hierarchie für den Morbi-RSA implementiert werden sollte, naturgemäß nicht neutral, sondern interessengeleitet. Sie werden diejenige Variante bevorzugen, die ihnen eine höhere Zuweisung aus dem Gesundheitsfond verspricht. In dem notwendigen, in Form von Anhörungen gepflegten Dialog zwischen dem Bundesversicherungsamt auf der einen und den Krankenkassen und ihren Verbänden auf der anderen Seite können die zusätzlich eröffneten Interpretationsspielräume fruchtbar genutzt werden.

1 Hintergrund und Aufgabenstellung

Im Klassifikationssystem für den morbiditätsorientierten Risikostrukturausgleich werden die Versicherten Morbiditätsgruppen zugeordnet, die in Hierarchien zusammengefasst sind. Dabei bestehen für einige Krankheitshierarchien unterschiedliche Möglichkeiten der Ausgestaltung, welche aus Sicht der medizinischen Klassifikation den gleichen Grad der Plausibilität aufweisen. Üblicherweise wird dann diejenige Ausgestaltungsvariante ausgewählt, welche bei der regressionsanalytischen Berechnung der Zuschläge mit einem höheren Anteil der erklärten Varianz (R^2) verbunden ist. Da von der Umgestaltung einer Hierarchie in der Regel jedoch nur vergleichsweise wenige Versicherte der Versichertenstichprobe betroffen sind, unterscheiden sich die zugehörigen R^2-Werte, die von der großen Zahl der „Nichtbetroffenen“ dominiert werden, vielfach erst in den hinteren Nachkommastellen. Die Entscheidung für die eine oder andere Variante auf der Basis der Differenz der R^2-Werte erscheint daher als wenig belastbar. Vor diesem Hintergrund war die Aufgabe des Gutachters, ein quantitatives Bewertungsverfahren zu entwickeln, das zu belastbareren Entscheidungen führt.

Im Rahmen einer Diskussion der geschilderten Problematik mit den Mitgliedern des Wissenschaftlichen Beirats zur Weiterentwicklung des Risikostrukturausgleichs und Mitarbeitern des Bundesversicherungsamtes wurde vom Gutachter das Grobkonzept eines Verfahrens vorgetragen, das im Folgenden als „Stichproben-Mikroskop-Design“ oder kurz „Mikroskop-Design“ bezeichnet wird. Dieses Konzept sollte am Beispiel einiger Ausgestaltungsvarianten ausgewählter Hierarchien verfeinert und erprobt werden.

Vom Bundesversicherungsamt wurden zu diesem Zweck je zwei Vergleiche von Ausgestaltungsvarianten der Hierarchien 02 (Neubildungen), 04 (Metabolische Erkrankungen) und 19 (Erkrankungen der Lunge) vorgeschlagen (zu Details s. Abschnitt 4).

Im empirischen Teil des Gutachtens wird das konzipierte Bewertungsverfahren an diesen insgesamt sechs Vergleichen im Detail erprobt und optimiert. Die hierfür erforderlichen Daten wurden einer Sonderauswertung des Bundesversicherungsamtes für das Gutachten entnommen.

2 Maßzahlen für die Güte der Anpassung im Regressionsmodell

2.1 Das klassische OLS-Regressionsmodell

Die Regressionsmethode wurde von Carl Friedrich Gauß ursprünglich als Ausgleichsverfahren entwickelt und in der Astronomie mit großem Erfolg angewendet. Die Schätzung der unter Zugrundelegung des Modells erwarteten Größen erfolgen mit der Methode der kleinsten Quadrate (Ordinary Least Square Estimation – OLS). Die Grundgleichung des inhomogenen Models lautet

(1) $y_i = ß_0 + ß_1 x_{1i} + ... + ß_k x_{ki} + e_i$.

Im Morbi-RSA sind die y_i die RSA-fähigen Ausgaben für den i-ten Versicherten und sie werden im Folgenden daher als a_i bezeichnet. Die Prädiktorvariablen x_j sind sämtlichst dichotome, nur der beiden Werte 0 und 1 fähige Größen (Indikatorvariable). Darüber hinaus gilt:

(2) $ß_0 = 0$ (homogenes Modell),

d. h. die Regressionshyperebene enthält den Ursprung des Koordinatensystems.

Viele der angenehmen Eigenschaften der Kleinste-Quadrate-Schätzungen gehen im homogenen Regressionsmodell im Allgemeinen verloren, sie gelten aber in dem vom Bundesversicherungsamt verwendeten homogenen Regressionsmodell weiterhin, weil die Versicherten dort eindeutig einer der 40 nach Alter und Geschlecht gebildeten Versichertengruppen zugeordnet sind und die Summe über die zugehörigen Indikatorvariablen stets 1 ergibt.

Die unter Zugrundelegung des Modells erwarteten (im Kontext die vorhergesagten oder „standardisierten“) Ausgaben berechnen sich nach Kalibrierung des Modells aus der Gleichung

(3) $\hat{a}_i = \hat{ß}_1 x_{1i} + ... + \hat{ß}_k x_{ki}$,

wobei die Schätzungen $\hat{ß}_j$ der Regressionskoeffizienten infolge der dichotomen Struktur der Prädiktoren die Zuschläge für die Versicherten darstellen.

Zu den grundlegenden Eigenschaften des Modells gehört, dass der Mittelwert der Ausgaben mit dem Mittelwert der standardisierten Ausgaben übereinstimmt und dass die Streuungszerlegung gilt:

(4) $$\frac{1}{n}\sum_{i=1}^{n}(a_i - \bar{a})^2 = \frac{1}{n}\sum_{i=1}^{n}(\hat{a}_i - \bar{\hat{a}})^2 + \frac{1}{n}\sum_{i=1}^{n}(a_i - \hat{a}_i)^2 ,$$

oder in Worten: Die Gesamtvarianz der Ausgaben setzt sich additiv aus zwei Komponenten zusammen, wobei die erste – die Varianz der standardisierten Ausgaben – von der Regression und die zweite (die Residualvarianz) durch Abweichung der tatsächlichen Ausgaben von den standardisierten verursacht wird.

2.2 Abweichungen des für den Morbi-RSA eingesetzten Verfahrens von der OLS-Regression

Das vom Bundesversicherungsamt zur Berechnung der Zuschläge eingesetzte Regressionsverfahren weicht vom klassischen OLS-Regressionsmodell in zwei Punkten ab.

1. Da die Ausgaben nicht ganzjährig Versicherter (mit Ausnahme der Verstorbenen) annualisiert werden, wird die Regression gewichtet vorgenommen. Dabei bekommt jeder Versicherte den Anteil des Jahres als Gewicht zugeordnet, in dem er in der gesetzlichen Krankenversicherung versichert war. Es handelt sich damit um ein Verallgemeinertes Kleinste-Quadrate-Modell (Generalized Least Square – GLS). Allerdings sind die Gewichte wegen der Dominanz der ganzjährigen Versicherten weitgehend identisch (das mittlere Gewicht beträgt 0,98), so dass die Lösungen in diesem speziellen GLS sich nur minimal von denjenigen aus dem OLS unterscheiden.
2. Darüber hinaus wird durch ein Iterationsverfahren sicher gestellt, dass keine negativen oder insignifikanten Zuschläge auftauchen. Wenn dieses im ersten Durchlauf passiert, werden die entsprechenden Zuschläge in einem erneuten Durchlauf auf Null beschränkt. Dieses Verfahren wird solange wiederholt, bis keine negativen Zuschläge mehr auftreten.

Infolge dieser Abweichungen vom OLS-Modell gelten einige Ausführungen des Abschnitts 2.1 und der Abschnitte 2.3 bis 2.6 nicht exakt, sondern nur in guter Näherung.

2.3 Das Bestimmtheitsmaß R^2

Dies ist die klassische Maßzahl zur Bestimmung des „Goodness of Fit“ im OLS-Regressionsmodell. Es beruht auf der Streuungszerlegung und wird zumeist in der Form

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (a_i - \hat{a}_i)^2}{\sum_{i=1}^{n} (a_i - \bar{a})^2} = 1 - \frac{\frac{1}{n}\sum_{i=1}^{n} (a_i - \hat{a}_i)^2}{\frac{1}{n}\sum_{i=1}^{n} (a_i - \bar{a})^2} \tag{5}$$

d. h. unter Verwendung des Anteils der Residualstreuung an der Gesamtstreuung berechnet.

Die Bezeichnung als erklärte Varianz wird aber erst dann verständlich, wenn man R^2 unter Verwendung der Streuungszerlegung umschreibt:

$$R^2 = \frac{\frac{1}{n}\sum_{i=1}^{n} (\hat{a}_i - \bar{\hat{a}})^2}{\frac{1}{n}\sum_{i=1}^{n} (a_i - \bar{a})^2} \quad . \tag{6}$$

Infolge der Streuungszerlegung kann der Zähler in (5) nicht größer werden als der Nenner und es gilt $0 \leq R^2 \leq 1$, wobei die Extreme (das Modell erklärt 0% bzw. 100% der Gesamtvarianz) in Regressionsmodellen für reale Daten praktisch nicht auftreten. Ferner bildet R^2 auch das Quadrat der multiplen Korrelation zwischen den Ausgaben und der Gesamtheit der Prädiktoren. Die multiple Korrelation ist definiert als die größtmögliche Korrelation, die zwischen den Ausgaben und einer Linearkombination der Prädiktoren berechnet werden kann. Angenommen wird das Maximum gerade für den Vektor $\hat{a} = (\hat{a}_1, \hat{a}_2, ..., \hat{a}_n)^T$ der standardisierten Ausgaben (T steht für die Operation der Matrixtransposition) und somit gilt:

(7) $$R^2 = (r_{a,\hat{a}})^2 \ .$$

Neben den vielen aufgezeigten positiven Eigenschaften hat R^2 allerdings eine schwerwiegende Schwäche, die zu Abschlägen bei der Bewertung der Eignung dieser Maßzahl insbesondere in Bezug auf die zu lösende Aufgabe führen muss: R^2 reagiert überaus empfindlich auf statistische Ausreißer, d. h. im Kontext: auf besonders teuere Versicherte. Die Darstellung (5) zeigt dabei, dass R^2 durch Ausreißer tendenziell herabgezogen wird. Der quadratische Term, der zu einem Ausreißer gehört, nimmt nämlich sowohl im Zähler als auch im Nenner einen wesentlich größeren Teil an der Gesamtsumme ein und die relative Abweichung der beiden Terme voneinander fällt viel kleiner aus als bei Betrachtung eines durchschnittlichen Versicherten. Dies sei an einem fiktiven einfachen Beispiel illustriert:

Angenommen 100 Versicherte weisen eine mittlere quadratische Abweichung vom Mittelwert in Höhe von 5.000 Euro^2 und einen mittleren Prädiktionsfehler in Höhe von 4.000 Euro^2 auf. Dann hat die Prädiktion ein R^2 in Höhe von R^2 = 1-4.000/5.000 = 0,200. Weiter angenommen die Ausgaben des 101. Versicherten liegen so weit über dem Durchschnitt, dass er eine quadratische Abweichung vom neu berechneten (leicht vergrößerten) Mittelwert von 50.000 Euro^2 und einen Prädiktionsfehler in Höhe von 42.000 Euro^2 aufweist. Vernachlässigt man die geringfügigen Änderungen der mittleren Quadrate der ersten 100 Versicherten, so ergibt sich nunmehr ein reduziertes R^2 in Höhe von R^2 = 1-46.000/55.000 = 0,164.

2.4 Maßzahlen, welche die Zahl der Prädiktoren berücksichtigen

Solche Maßzahlen wurden im Zusammenhang mit den früher sehr beliebten (heute nicht mehr empfohlenen)[1)] Verfahren der schrittweisen Variablenselektion eingeführt und verwendet, da ein Vergleich der R^2-Werte von zwei Modellen mit unterschiedlicher Variablenzahl in die Irre führen kann. Jede Zunahme eines weiteren Prädiktors erhöht R^2 und zwar unabhängig davon, ob der Prädiktor einen Beitrag zur Erklärung leistet oder nicht. Dies folgt aus der Darstellung (6). In den Zähler des Bruchs in (6) sind nämlich

1 Vgl. Harrel, Jr. (2001), S. 56 ff.

ausschließlich modellerzeugte Werte involviert und in den Nenner ausschließlich Beobachtungswerte. Es lässt sich zeigen, dass der Zähler bei Hinzunahme eines weiteren Prädiktors wächst, während der Nenner naturgemäß davon unberührt bleibt. Im Ergebnis wird R^2 bei Hinzunahme eines weiteren Prädiktors größer.

2.4.1 Das adjustierte R^2

Zu dieser Maßzahl kommt man in natürlicher Weise, wenn berücksichtigt wird, dass Zähler und Nenner in (4) keine erwartungstreuen Schätzer der Residual- bzw. der Gesamtvarianz darstellen. Ersetzt man sie durch die auf Erwartungstreue korrigierten Schätzungen, so erhält man das von Henri Theil vorgeschlagene adjustierte R^2 (vgl. Theil 1971):

$$(8) \qquad R^2_{adj} = 1-(1-R^2)\frac{n-1}{n-p-1} = R^2-(1-R^2)\frac{p}{n-p-1}$$

In der zweiten Darstellung des adjustierten Bestimmtheitsmaßes in (8) stellt der abgezogene Term (der sog. Strafterm) eine Korrektur dar, die das Anwachsen des Bestimmtheitsmaßes mit der Zahl p der Parameter berücksichtigt.

2.4.2 Mallow's C_p

Sei $RSS(p) = \sum_{i=1}^{n}(a_i - \hat{a}_i)^2$ die Summe der quadrierten Residuen bei Kalibrierung des Modells mit einer Auswahl von p Prädiktoren. Dann wird Mallow's C_p folgendermaßen berechnet:

$$(9) \qquad C_p = \frac{RSS(p)}{\frac{1}{n}\sum_{i=1}^{n}(a_i - \hat{a}_i)^2} - n + 2p\,,$$

wobei der Nenner aus dem Modell mit allen Prädiktoren berechnet wird. Außer in extremen Situationen ist der Erwartungswert von Cp $\geq$ p und liegt nahe p. Wenn das C_p eines Modells A näher an p liegt, als das C_p eines Modells B, wird A als das bessere Modell angesehen.

2.4.3 Informationskriterien

Informationskriterien sind Kriterien zur Modellauswahl, welche die Anpassungsgüte des Modells und seine Komplexität – gemessen an der Anzahl p der Parameter – berücksichtigen. Sie verfolgen einen informationstheoretischen Ansatz und finden auch in Klassifikationsverfahren Anwendung. Da sie auf der Varianz der Residuen beruhen, schneidet ein Modell umso besser ab, je kleiner der Wert des verwendeten Informationskriteriums ausfällt.

Das älteste wurde von Akaike im Jahr 1972 vorgeschlagen und berechnet sich im Kontext der OLS-Regression wie folgt (vgl. Greene 2008, S. 143):

(10) $$AIC = \ln(\frac{1}{n}\sum_{i=1}^{n}(a_i - \hat{a}_i)^2) + \frac{2p}{n}$$

Die Anzahl der Parameter wird dabei „strafend“ berücksichtigt, um die Bevorzugung umfassender Modelle mit vielen Parametern zu vermeiden. Allerdings wird der Strafterm mit wachsendem Stichprobenumfang n schnell sehr klein. Schwarz hat 1978 ein Bayessches Informationskriterium vorgeschlagen, bei dem der Strafterm bei wachsendem n nicht so schnell schrumpft. Dieses genügt im OLS-Regressionsmodel der folgenden Gleichung (vgl. Greene 2008, S. 143):

(11) $$BIC = \ln(\frac{1}{n}\sum_{i=1}^{n}(a_i - \hat{a}_i)^2) + \frac{p}{n}\ln(n)$$

Wenn es um die Wahl von AIC oder BIC geht, so findet BIC häufige Anwendung in der Soziologie und AIC wird von den Ökonometrikern bevorzugt (Kuha 2004).

Die Bezeichnung "Strafterm" stammt von keinem der beiden Erfinder, sondern ist erst später von den Anwendern der Informationskriterien, insbesondere im OLS-Regressionsmodel, eingeführt worden. Die Informationskriterien wurden jedoch für andere Fragestellungen entwickelt und ihre Logik kann sich im Rahmen des OSL-Schätzverfahrens nicht erschließen.

Obwohl sehr ähnlich aufgebaut, liegen den beiden Informationskriterien unterschiedliche Konzepte zugrunde. Zwar betrachten sowohl Akaike als auch Schwarz eine Verteilungsfamilie mit den Dichten $f(x|\theta)$, wobei der Vektor $\theta=(\theta_1, \theta_2, \ldots. \theta_p)^T$ die p unbekannten Parameter zusammenfasst. Zur Anpassung des Modells an den Beobachtungsvektor $x=(x_1,x_2,\ldots..,x_n)^T$ verfolgt Akaki allerdings den Ansatz, die Maximum-Liklehood-Methode (ML-Methode) auf die Aufgabenstellung der Modellauswahl mit variablem p zu erweitern, und Schwarz den Ansatz einer Bayes-Schätzung des Parametervektors. Beide suchen eine Verteilung, die gut zu den Daten passt, aber ihre jeweilige Definition von "guter" Anpassung unterscheidet sich fundamental. Während es Schwarz darum geht, dasjenige Modell zu finden, welches mit der höchsten Wahrscheinlichkeit das wahre Model ist, das den Daten zugrunde liegt, schließt Akaike ausdrücklich aus, dass das wahre Modell in der betrachteten Verteilungsfamilie liegt, und sucht nach derjenigen Verteilung der Familie, welche die wahre Verteilung am besten approximiert. Der Abstand zwischen der approximierenden und der wahren Verteilung, gemessen in einer eng mit der Maximum-Liklehood-Schätzmethode verbunden Metrik, wird von Akaike als eine Art Vorhersagefehler aufgefasst, und der Strafterm des AIC geht hervor aus einer Korrektur auf Erwartungstreue bei Schätzung dieses Fehlers (zu Details s. Akaike 1974).

Im Rahmen der Bayes-Schätztheorie wird von einer a-priori-Verteilung über den unbekannten zu schätzenden Parameter ausgegangen und dann mit Hilfe der bedingten Verteilung $f(x|\theta)$ des Beobachtungsvektors x, gegeben der Parameter θ, die a-posteriori-Verteilung $f(\theta|x)$ für θ, gegeben x, nach dem Satz von Bayes berechnet. Der Erwartungswert der a-posteriori-Verteilung wird als Bayes-Schätzung von θ bezeichnet.

Schwarz ging bei der Konstruktion des BIC von der bekannten Möglichkeit aus, den ML-Schätzer von θ bei vorgegebener Dimension p für $n \to \infty$ als führenden Term einer asymptotischen Entwicklung des Bayes-Schätzers erhalten zu können, und zwar unabhängig von der speziellen a-priori-Verteilung (sofern deren Dichte überall größer als Null ist). Um einen ähnliche Grenzwertsatz auch bei variablem p zu erhalten, betrachtet er eine spezielle Klasse von a-priori-Verteilungen, deren Dichte sich als gewichtete Summe darstellen lässt. Dabei werden die Dichten der zur Konkurrenz zugelassenen Modelle M_j (mit unterschiedlicher Variablenzahl p) gewichtet mit der jeweiligen a-priori-Wahrscheinlichkeit, dass M_j das wahre Model ist.

Die asymptotische Entwicklung der Bayes-Schätzung führt wieder unabhängig von der speziellen a-priori-Verteilung mit der genannten Struktur auf den ML-Schätzer als führenden Term. Der Strafterm des BIC ergibt sich dann als nächster Term der Entwicklung (Approximationsterm zweiter Ordnung).

Eine ausführliche Analyse (die den Rahmen des vorliegenden Gutachtens sprengt), auf welche Weise sich die Bestrafung von Komplexität eines Models bei der Auswahl von Modellen nach Maßgabe von BIC (und auch nach Maßgabe von AIC) ergibt, findet sich bei Kuha (2004).

Die Vor- und Nachteile der beiden erörterten Informationskriterien wurden von verschieden Autoren im Vergleich diskutiert. Für große Stichproben ist BIC besser geeignet als AIC, da BIC als Selektionskriterium konsistent ist. Dies bedeutet, dass die Wahrscheinlichkeit, mit der das wahre Model M_w auf der Basis von BIC gewählt wird, für $n \to \infty$ gegen Eins geht, sofern M_w in der untersuchten Verteilungsfamilie enthalten ist. Das gilt nicht für AIC, das für $n \to \infty$ dazu tendiert, immer komplexere Modelle auszuwählen. (s. z. B. Hastie et al., S. 208). Auf der anderen Seite ist, wie Kuha (2004) ausführt, das Argument der Konsistenz nur dann überzeugend, wenn man die Existenz eines wahren Models voraussetzt, was Akaike ja ausdrücklich nicht tut. Sein Ansatz misst die Güte eines Models an der Fähigkeit der Vorhersage. AIC (aber nicht BIC) ist nach Kuha (2004) vor diesem Hintergrund in einem gewissen Sinn asymptotisch effizient, da in Fällen, in denen das wahre Model von unendlicher Dimension ist, oder seine Dimension mit dem Stichprobenumfang wächst, der mittlere quadratische Vorhersagefehler bei Modelauswahl nach Maßgabe von AIC asymptotisch für $n \to \infty$ der kleinst mögliche ist. Kuha (2004) empfiehlt daher immer beide Informationskriterien zu verwenden, und, wenn sie zu unterschiedlichen Ergebnissen führen, weitere Kriterien heranzuziehen.

Da es keinen Grund zu der Annahme gibt, dass die Differenz der Prädiktorenzahl der zu bewertenden Hierarchievarianten im Klassifikationssystem des Morbi-RSA mit wachsendem Stichprobenumfang zunimmt, dürfte BIC für die im vorliegenden Gutachten erörterte Bewertungsaufgabe das geeignetere Informationskriterium für die Güte der Anpassung sein, falls sich die Hierarchievarianten in der Zahl der Prädiktoren unterscheiden.

2.5 Das Predictive Ratio

Das Vorhersageverhältnis (so der deutsche Name für diese Maßzahl, im Kontext des Risikostrukturausgleichs häufig auch als Deckungsquote bezeichnet) wird bei Anwendungen des Regressionsmodells für Zwecke der Risikoadjustierung häufig verwendet. Dabei wird es für Versichertengruppen berechnet, die entweder nach Höhe der Ausgaben oder in Bezug auf das Vorliegen bestimmter Diagnosen definiert sind (z. B: Versicherte, deren Ausgaben im obersten Quintil liegen oder Versicherte mit koronarer Herzkrankheit, mit psychiatrischen Diagnosen o. ä.).

Seine Ableitung folgt einem einfachen Gedanken. Die Summe der vorhergesagten Aussagen der Versicherten der betrachteten Gruppe im Zähler wird ins Verhältnis gesetzt zur Summe der tatsächlichen Ausgaben dieser Versicherten. Wenn die betrachtete Versichertengruppe K Versicherte enthält und wir diese von 1 bis K durchnummerieren, so gilt also:

$$PR = \frac{\sum_{i=1}^{K} \hat{a}_i}{\sum_{i=1}^{K} a_i} \tag{12}$$

Bedauerlicherweise kann diese Maßzahl zur Lösung des geschilderten Problems nicht herangezogen werden. Für Versichertengruppen, deren zugehörige Indikatorvariablen als Prädiktoren im Regressionsmodell aufgenommen sind, gilt nämlich

$$\sum_{i=1}^{K} \hat{a}_i = \sum_{i=1}^{K} a_i \quad . \tag{13}$$

Es ist also PR=1 für solche Gruppen und dies gilt auch für Zusammenfassungen (Vereinigungsmengen) solcher Versichertengruppen, weil sich die PR einer Vereinigungsmenge als gewichteter Mittelwert aus den PRs der zusammengefassten Versichertengruppen berechnet.

2.6 Maßzahlen, in denen das Quadrat durch den Absolutbetrag ersetzt wird

2.6.1 Mean Absolute Prediction Error (MAPE)

Diese Maßzahl ist der Residualvarianz ähnlich, verwendet aber anstelle des Quadrats den absoluten Betrag und ist daher weniger anfällig gegenüber statistischen Ausreißern:

$$MAPE = \frac{\sum_{i=1}^{n} \left| a_i - \hat{a}_i \right|}{n} \tag{14}$$

Der mittlere absolute Vorhersagefehler wird im Regressionsmodell für den Morbi-RSA in Euro ausgewiesen und ist daher eine sehr transparente und gut interpretierbare Maßzahl. Ein Modell ist umso besser, je kleiner der Wert von MAPE ausfällt. Da in die Berechnung von MAPE Beobachtungswerte und modellerzeugte Werte einfließen, kann MAPE bei Hinzunahme eines zusätzlichen Prädiktors sowohl größer als auch kleiner werden.

2.6.2 Cumming's Predicton Measure (CPM)

Cumming und Cameron (2002) haben – gefördert von der Society of Actuaries – die in den USA angewendeten Versichertenklassifikationsverfahren und Regressionsmodelle zur Risikoadjustierung hinsichtlich ihrer prädiktiven Qualität miteinander verglichen. Neben R^2 und MAPE haben sie zu diesem Zweck eine weitere von Cumming vorgeschlagene Maßzahl verwendet, die analog zu R^2 aufgebaut ist, aber das Quadrat durch den Absolutbetrag ersetzt:

$$CPM = 1 - \frac{\sum_{i=1}^{n} |a_i - \hat{a}_i|}{\sum_{i=1}^{n} |a_i - \bar{a}_i|} \tag{15}$$

Wenn man den abgezogenen Bruch um den Faktor $1/n$ erweitert, so steht im Zähler des Bruches MAPE und im Nenner das Analogon zur Gesamtvarianz, das als Mean Absolute Deviation (MAD) bezeichnet wird. Während MAPE ein absolutes Maß darstellt, ist CPM ein relatives Maß, das nach Ansicht von Cumming zwischen 0 und 1 liegt und den Anteil der erklärten Variation anzeigt, wobei die extremen Werte bedeuten, dass das Modell 0% bzw. 100% der Ausgabenvariation „erklärt". In diesem Punkt irrt sich Cumming jedoch, da es eine (4) analoge „Variationszerlegung" für die MAD nicht gibt, kann der Zähler im Bruch von (15) theoretisch den Nenner übersteigen und CPM wird negativ (was in extremen Fällen auch beobachtet werden kann). Aus dem gleichen Grund kann CPM bei Hinzunahme eines zusätzlichen Prädiktors sowohl größer, als auch kleiner werden. Eine Adjustierung von CPM hinsichtlich der Zahl der im Modell inkorporierten Prädiktoren ist daher nicht nur nicht erforderlich, sondern muss sogar als falsch angesehen werden. Trotz dieser Eigenschaften ist CPM für Modellvergleiche eine interessante Maßzahl, da sie nicht die Empfindlichkeit von R^2 gegenüber teuren Versicherten ausweist.

2.6.3 Alternative Prediction Measure (APM)

Wegen des Fehlens eines Analogons für das MAD fallen (5) und (6) auseinander und liefern, wenn man das Quadrat durch den Absolutbetrag ersetzt, unterschiedliche Maßzahlen, so dass das CPM nicht übereinstimmt mit

$$APM = \frac{\frac{1}{n}\sum_{i=1}^{n} |\hat{a}_i - \bar{\hat{a}}|}{\frac{1}{n}\sum_{i=1}^{n} |a_i - \bar{a}|} \tag{16}$$

APM wächst wie R^2 mit der Zahl der Prädiktoren und müsste geeignet adjustiert werden. Allerdings erscheint CPM a priori besser geeignet als APM, da in den Zähler von (16) ausschließlich vorhergesagte Ausgaben involviert sind, während im Zähler von (15) die vorhergesagten mit den tatsächlichen Ausgaben verglichen werden.

3 Untersuchungsansatz: Mikroskop-Design und Resampling

3.1 Mikroskop-Design

Um die Verdünnung infolge einer großen Zahl von Versicherten, die von den Unterschieden in der Ausgestaltung einer Hierarchie nicht betroffen sind, entgegen zu wirken, und die Differenzen der herangezogenen Maßzahlen gleichsam unter dem Mikroskop betrachten zu können, wird der folgende Ansatz gewählt:

Bezogen auf eine bestimmte vorgegebene Hierarchie mit zwei verschiedenen Varianten wird aus der Menge der vom Unterschied nichtbetroffenen Versicherten eine Unterstichprobe (ohne Zurücklegen) gezogen und zusammen mit der Gesamtheit der Versicherten der Betroffenen ausgewertet, so dass das Design in Richtung eines balancierten Designs verändert wird. Im empirischen Teil des Gutachtens wurden dabei zwei verschiedene Operationalisierungen des Begriffes „Betroffene“ erprobt (eine enge und eine weiter gefasste).

Bezeichnet N_{NB} den Stichprobenumfang der Nichtbetroffenen (also der Unterstichprobe) und N_B die Zahl der Betroffenen, so wurde für Optimierungszwecke im empirischen Teil des Gutachtens

(17) $N_{NB} = qN_B$ mit q=1, 2 oder 3

erprobt. Allerdings sollte N_{NB} nicht unter 2000 fallen und wurde ggf. entsprechend gesetzt und q=2 und q=3 war nur für die enge Operationalisierung von „betroffen“ anzuwenden.

Die Unterstichprobe, zusammen mit den betroffenen Versicherten, bildet dann einen Datensatz, der als „Bewertungsstichprobe“ bezeichnet wird und den Stichprobenumfang

(18) $m = N_{NB} + N_B$

besitzt.

Es schien allerdings nicht zweckmäßig, das Modell aus der Bewertungsstichprobe neu zu kalibrieren, da die Schätzung der über 150 Regressionskoeffizienten bei einem kleinen m durch riesige Varianzen und damit einhergehenden Instabilitäten gefährdet wäre.

Für jede der beiden Ausgestaltungsvarianten der betrachteten Hierarchie lassen sich aber auch ohne Neukalibrierung des Modells aus der Bewertungsstichprobe alle in Abschnitt 2 diskutierten Maßzahlen berechnen. Die Regressionskoeffizienten stammen dabei aus der Kalibrierung des Modells in der jeweiligen Variante an der vollen Versichertenstichprobe. Das hat allerdings zur Folge, dass die Mittelwerte der Ausgaben und der standardisierten Ausgaben, berechnet aus der Bewertungsstichprobe, im Allgemeinen nicht mehr übereinstimmen und die Streuungszerlegung (4) nicht mehr gilt. Daher werden sich auch die Werte von R^2, je nachdem ob man die Formel (5), (6) oder (7) heranzieht im Allgemeinen voneinander unterscheiden und R^2 kann auch negativ werden.

3.2 Zur Erprobung herangezogene Maßzahlen

Es ist nicht zweckmäßig, alle in Abschnitt 2 diskutierten Maßzahlen im empirischen Teil in gleicher Tiefe zu erproben. Einige wurden schon aus grundsätzlichen Erwägungen ausgeschlossen. Für andere wiederum ist a priori klar, dass es jeweils Gruppen unter ihnen gibt, deren Mitglieder zwangsläufig alle zum gleichen Ergebnis in der Bewertung führen. Der Grund hierfür ist, dass die Nenner in (5) bzw. (6) und (15) bzw. (16) in allen Ausgestaltungsvarianten einer Hierarchie gleich ausfallen.[2] In die Nenner sind ja nur die tatsächlichen Ausgaben der Versicherten der Bewertungsstichprobe involviert und nicht die vorhergesagten. Es würde daher genügen, sich auf die Zähler zu beschränken, also z. B. MAPE zu betrachten (anstelle von CPM) oder die Residualvarianz bzw. das darauf basierende BIC zu betrachten und nicht R^2.[3] Allerdings wäre es nicht zweckmäßig R^2, die üblicherweise (und bisher einzige) verwendete Maßzahl, aus dem Mikroskop-Design auszuschließen. Darüber hinaus zieht man zur Frage, wie groß q gewählt werden sollte und wie viele Wiederholungen vorgesehen werden sollten, besser die komplexer gebauten Verhältniszahlen R^2 und CPM heran, als sich gerade für diese Untersuchungsteile auf die einfacheren und größeren Zählerinformationen zu beschränken. Zu beachten ist auch, dass die zu untersuchenden Ausgestaltungsvarianten der Hierarchien „Erkrankungen der Lunge“ und „Neubildungen“ sich jeweils nicht in der Zahl der Prädiktoren unterscheiden (es gilt p=153 bzw. p=152 für alle Varianten, vgl. Abschnitt 4). Nur in der Hierarchie „Metabolische Erkrankungen“ variiert p von p=152 über p=153 bis zu p= 154 (vgl. Abschnitt 4). Daher kann in den beiden zuerst genannten Hierarchien auf ein Informationskriterium als zusätzliche Maßzahl verzichtet werden (wenn R^2 herangezogen wird) und ein solches wird nur für die Vergleiche der zuletzt genannten Hierarchie benötigt.

Vor diesem Hintergrund wurden im empirischen Teil des Gutachtens folgende Maßzahlen verwendet:

- Für Zwecke der Optimierung des Designs: R^2 und CPM, sowie ggf. BIC
- Für die Erprobung des optimierten Mikroskop-Designs an ausgewählten Beispielen: R^2, MAPE und $r = r_{a,\hat{a}}$, sowie ggf. BIC

Dabei wurde die Korrelation zwischen den tatsächlichen und den vorhergesagten Ausgaben als Maßzahl hinzu genommen, weil sie einem vertrauten statistischen Konzept folgt, gut interpretierbar ist und weil ihr Quadrat (wie oben ausgeführt) in der Bewertungsstichprobe nicht exakt mit R^2 übereinstimmt. Die Maßzahl BIC wurde nur für Vergleiche der Hierarchie „Metabolische Erkrankungen“ berechnet.

---

[2] Das ist beim Vergleich von Klassifikationssystemen natürlich nicht der Fall, weshalb Cumming das relative Maß CPM dem absoluten Maß MAPE vorzog.

[3] Dies Argument verliert an Gültigkeit, wenn man die Konsistenz der Entscheidungen bei der Ausgestaltung einer Hierarchie über mehrere Jahre prüfen will. Für solche Zwecke sollte MAPE stets durch CPM flankiert werden.

3.3 Resampling

Betrachtet man die folgenden Differenzen

(19) $$D_1 = R_2^2 - R_1^2, \quad D_2 = CPM_2 - CPM_1, \quad D_3 = MAPE_2 - MAPE_1, \quad D_4 = r_2 - r_1, \quad D_5 = BIC_2 - BIC_1$$

wobei 2 und 1 auf die jeweilige Ausgestaltungsvariante der betrachteten Hierarchie verweist, so soll sich zur Beurteilung der Stabilität des Verfahrens die Bewertung auf die Verteilungen dieser fünf Differenzen stützen (D_5 ist nur für Vergleiche der Hierarchie „Metabolische Erkrankungen“ von Interesse, s. o.).

Da diese Verteilungen nicht bekannt sind, müssen sie geschätzt werden, wobei im Folgenden ein Ansatz verfolgt wird, der dem in der Statistik verwendeten Begriff des Resampling subsumiert werden kann. Hierzu ist die Ziehung der beschriebenen Unterstichprobe aus den Nichtbetroffenen, die ja ihrerseits ohne Zurücklegen gezogen wurde, nun n-mal mit Zurücklegen zu wiederholen (d. h. Versicherte, die in vorangegangenen Ziehungen in die Unterstichprobe gelangt sind, können wieder gezogen werden). Auf diese Weise entstehen n Bewertungsstichproben mit jeweils n Ausprägungen für die zwei zur Bewertung herangezogenen Differenzen. Deren empirischen Verteilungen können dann z. B. durch Darstellung von Histogrammen oder Berechnung von Mittelwerten und Perzentilen ausgewertet werden.

4 Zur Erprobung ausgewählte Hierarchien und Ausgestaltungsvarianten

Die Auswahl der nachfolgend analysierten Hierarchien und Ausgestaltungsvarianten wurde vom Bundesversicherungsamt vorgenommen, das auch die Ziehung der Bewertungsstichproben und die Berechnung der Maßzahlen übernommen hat. Die ausgewählten Beispiele beziehen sich auf Entscheidungsprobleme, die im Rahmen der Festlegungen zum Klassifikationssystem für den Jahresausgleich 2011 relevant waren und in den Erläuterungen zur Festlegung dargestellt werden.

4.1 Erkrankungen der Lunge

In dieser Hierarchie wurde ein Ausgangsmodell mit zwei Ausgestaltungsvarianten verglichen

4.1.1 Variante 1: Ausgangsmodell

Im Ausgangsmodell wird die neu in die Krankheitsauswahl aufgenommene DxG454 (Bronchiektasen) als eigenständige, nicht hierarchisierte Risikogruppe in das Modell aufgenommen (vgl. Abbildung 4-1). Da diese Ausgestaltung bei der Modellanpassung aus normativen Gründen verworfen wurde, wurde die Variante im Festlegungsentwurf nicht dokumentiert.

4.1.2 Variante 2: Einbindung der DxG454 (Bronchiektasen)

Die DxG 454 („Bronchiektasen“) wird als eigenständige Zuschlagsgruppe aufgenommen, aber in der Hierarchie zwischen den HMG108 und HMG109 eingeordnet (vgl. Abbildung 4-2).

4.1.3 Variante 3: Aufteilung der HMG107 (Mukoviszidose)

Die DxG454 wird (wie im Ausgangsmodell) als eigenständiger Risikofaktor (ohne Hierarchisierung) im Regressions- und Zuweisungsverfahren berücksichtigt. Die Änderung zum Ausgangsmodell ergibt sich wie folgt:

- Die (H)MG107 wird altersabhängig gesplittet. Es entstehen die beiden neuen
  - HMG193 (Mukoviszidose < 12 Jahre) und
  - HMG194 (Mukoviszidose ≥ 12 Jahre)
- Für eine Gruppierung in die HMG194 muss zudem eine spezifische medikamentöse Therapie vorliegen.

Die Anpassungen entsprechen (mit Ausnahme der Berücksichtigung der DxG454 – Bronchiektasen) dem „Modell 4“ in Abschnitt 18.4. der Dokumentation zum Festlegungsentwurf (vgl. Abbildung 4-3).

4.1.4 Definition von „betroffen“ und realisierte Stichprobenumfänge

In der weiten Definition werden alle diejenigen Versicherte als betroffen eingestuft („Betroffene I“), denen mindestens eine der nachfolgenden Morbiditätsgruppen zugeordnet ist: (H)MG107, (H)MG108, (H)MG109, (H)MG110, (H)MG111, (H)MG112 und DxG454.

In der engeren Definition in Bezug auf die Variante 2 werden nur Versicherte als betroffen eingestuft, bei denen Bronchiektasen (DxG454) vorliegen („Betroffene II“).

In der engeren Definition in Bezug auf die Variante 3 werden nur Versicherte als betroffen eingestuft, die im Ausgangsmodell die MG107 (Mukoviszidose) aufgewiesen haben („Betroffene III“).

Die folgende Tabelle vermittelt einen Überblick über die verschiedenen Stichprobenumfänge, wobei die drei Varianten mit V 1, V 2 und V 3 abgekürzt werden. Zum Vergleich: der vollständige Datensatz umfasst 4.428.698 Pseudonyme, d. h. Versicherte.

Tabelle 4-1: Stichprobenumfänge m der Bewertungsstichproben für die Vergleiche in der Hierarchie „Erkrankungen der Lunge“ *)

	q = 1			q = 2		q = 3	
Vergleich	N_B	N_{NB}	m	N_{NB}	m	N_{NB}	m
V 1 mit V 2, Betroffene I	220.934	220.934	441.868				
V 1 mit V 2, Betroffene II	1.753	2.000	3.753	4.000	5.753	6.000	7.753
V 1 mit V 3, Betroffene I	220.934	220.934	441.868				
V 1 mit V 3, Betroffene III	462	2.000	2.462	4.000	4.462	6.000	6.462

*) N_B = Zahl der Betroffenen, N_{NB} = Umfang der Stichprobe aus den Nichtbetroffenen
m = Umfang der Bewertungsstichprobe

Die Größe von N_{NB} wurde dabei folgendermaßen festgelegt:

- $N_{NB} = q*NB$ im Fall, dass $N_B >= 2.000$ gilt.
- $N_{NB} = q*2.000$ im Fall. dass $N_B < 2.000$ gilt.

Was die Zahl p der Prädiktoren anbetrifft, so gilt p=153 für alle drei Varianten.

Abbildung 4-1 Hierarchie „Erkrankungen der Lunge“ im Ausgangsmodell

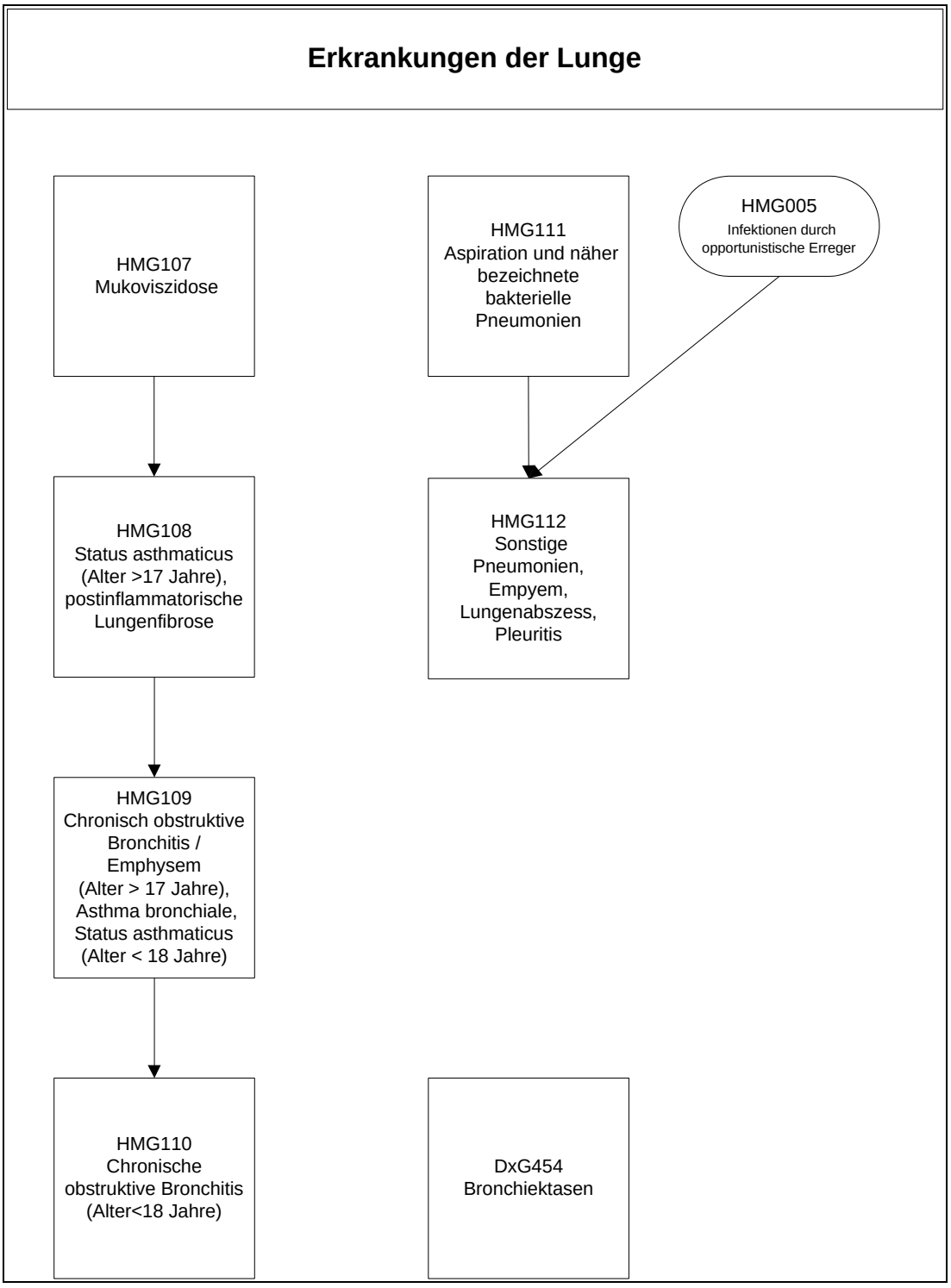


Abbildung 4-2: Geänderte Berücksichtigung der DxG454 (Bronchiektasen) in der Variante 2

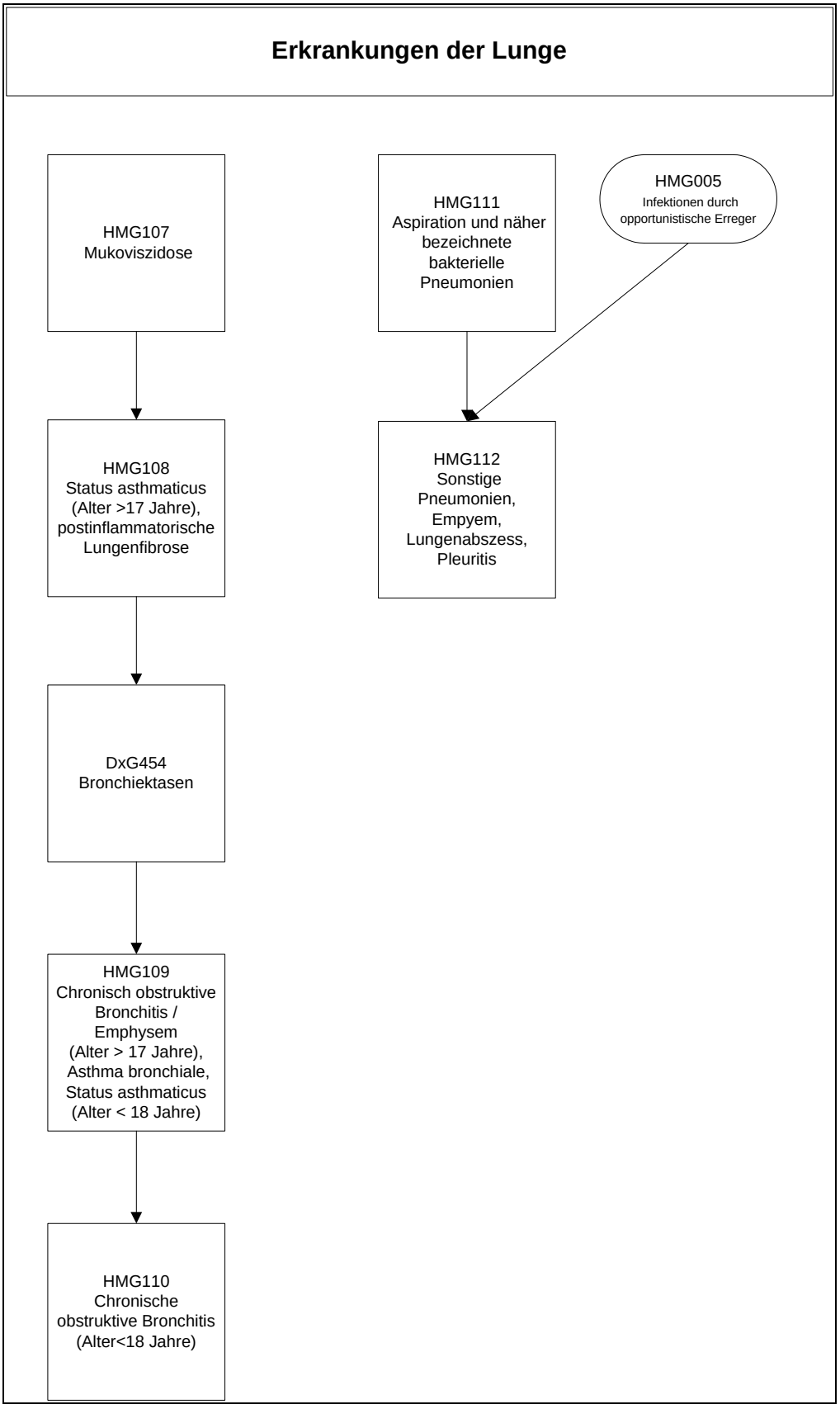
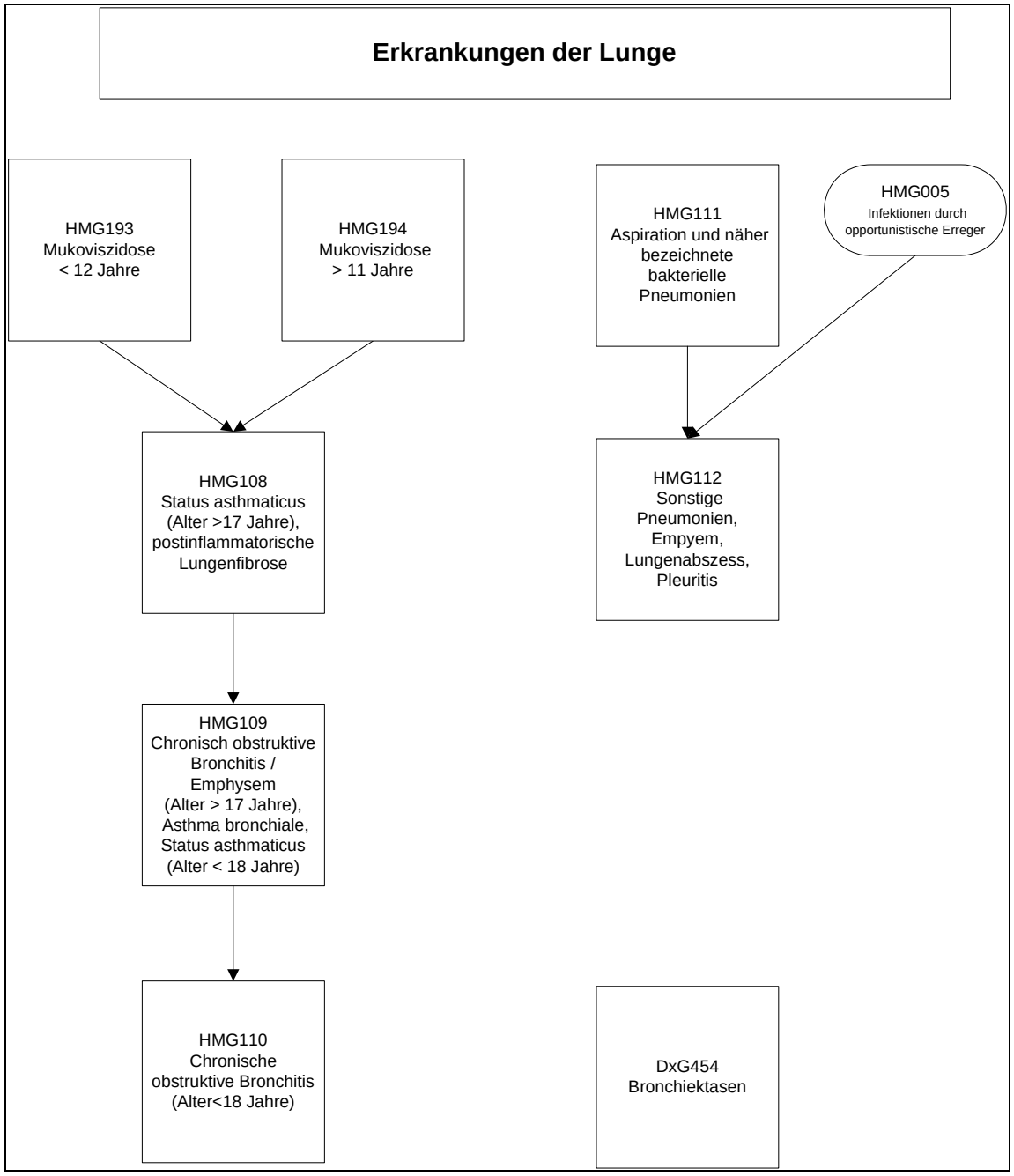


Abbildung 4-3: Geänderte Einbindung der Mukoviszidose in die Hierarchie in der Variante 3



4.2 Neubildungen

In dieser Hierarchie wurde zunächst das Ausgangsmodell mit einer Ausgestaltungsvariante verglichen. Im zweiten Vergleich werden zwei Ausgestaltungsvarianten (ohne Bezug auf das Ausgangsmodell) miteinander verglichen.

4.2.1 Variante 1: Modell 0 der Erläuterung zur Festlegung (Ausgangsmodell)

Ausgangsmodell der Betrachtung ist der Status quo des Klassifikationsmodells 2010 vor Aufnahme neuer ICD in das Modell (Modell 0 der Erläuterung zur Festlegung von Morbiditätsgruppen, Zuordnungsalgorithmus, Regressionsverfahren und Berechnungsverfahren, S. 53, Tabelle 4), vgl. Abbildung 4-4.

4.2.2 Variante 2: Modell A.4 in der Erläuterung zur Festlegung

Variante 2 stellt eine Modifikation des Anhörungsvorschlags dar: Das Modell ist allerdings schon bereinigt um die zunächst vorgesehene, jedoch aufgrund der Erkenntnisse des Anhörungsverfahrens aus fachlichen Gründen nicht umgesetzte Nutzung der Zusatzinformationen „Chemo- und Strahlentherapie“. Im Vergleich zum Status quo wird die Diagnose D47.1 von HMG014 in HMG007 verschoben. Zusätzlich wird die im Rahmen der Anhörung vorgeschlagene Verschiebung der ICD-Codes C93.0- und C94.0- von HMG006 in HMG004 geprüft. Variante 2 entspricht somit dem Modell A.4 der Erläuterungen zur Festlegung (S. 53, Tabelle 4).

4.2.3 Variante 3: Modell A.3 der Erläuterungen zur Festlegung

Variante 3 entspricht, bis auf die noch nicht geprüfte Zuordnung des ICD-Codes C94.0-, der Variante 2. In Variante 3 wird in Abweichung zu Variante 2 der ICD C94.0- in HMG006 belassen. Variante 3 entspricht somit Modell A.3 der Erläuterungen zur Festlegung (Modell A.3 der Erläuterung zur Festlegung von Morbiditätsgruppen, Zuordnungsalgorithmus, Regressionsverfahren und Berechnungsverfahren, S. 53, Tabelle 4).

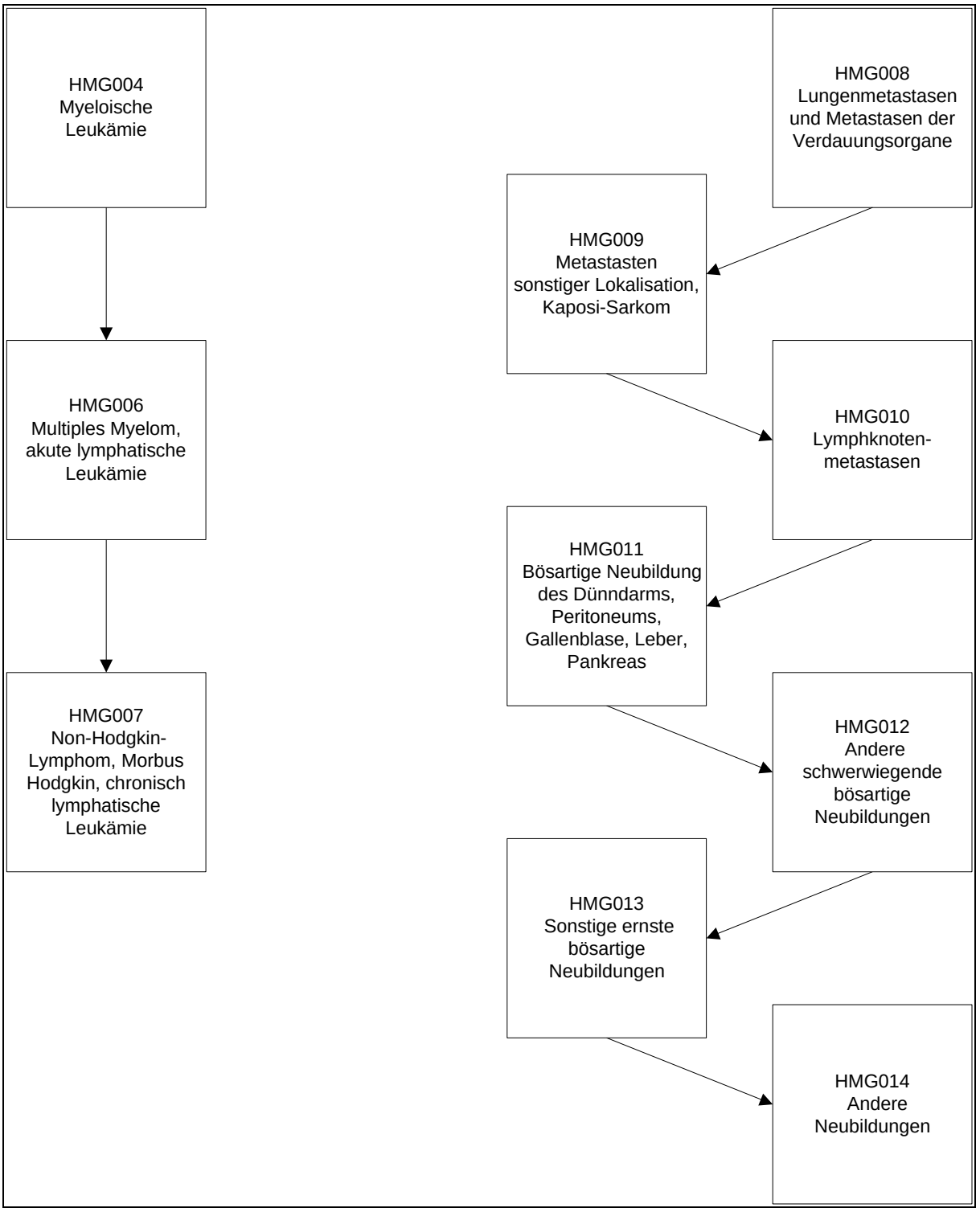
4.2.4 Definition von „betroffen“ und realisierte Stichprobenumfänge

In der weiten Definition werden alle diejenigen Versicherte als betroffen eingestuft („Betroffene I“), denen mindestens eine HMGs der Hierarchie „Neubildungen“ zugeordnet ist.

In der engeren Definition in Bezug auf die Variante 2 werden nur Versicherte als betroffen eingestuft („Betroffene II“), die von der Verschiebung der Diagnosen D47.1, C93.0- und /oder C94.0- betroffen sind (Veränderung Variante 1 zu Variante 2).

In der engeren Definition in Bezug auf die Variante 3 („Betroffene III“) werden nur Versicherte als betroffen eingestuft, die vom Wegfall der Diagnose C94.0- betroffen sind (Veränderung Variante 2 zu Variante 3).

Abbildung 4-4: Hierarchie „Neubildungen“ im Status quo



Die folgende Tabelle weist die verschiedenen Stichprobenumfänge aus.

Tabelle 4-2: Stichprobenumfänge der Bewertungsstichproben für die Vergleiche in der Hierarchie der Neubildungen*)

	q = 1			q = 2		q = 3	
Vergleich	N_B	N_{NB}	m	N_{NB}	m	N_{NB}	m
V 1 mit V 2, Betroffene I	151.183	151.183	302.366				
V 1 mit V 2, Betroffene II	1.207	2.000	3.207	4.000	5.207	6.000	7.207
V 2 mit V 3, Betroffene I	151.183	151.183	302.366				
V 2 mit V 3, Betroffene III	47	2.000	2.047	4.000	4.047	6.000	6.047

*) N_B = Zahl der Betroffenen, N_{NB} = Umfang der Stichprobe aus den Nichtbetroffenen
m = Umfang der Bewertungsstichprobe

Was die Zahl p der Prädiktoren anbetrifft, so gilt p=152 für alle drei Varianten.

4.3 Metabolische Erkrankungen

Auch in dieser Hierarchie wurde zunächst das Ausgangsmodell mit einer Ausgestaltungsvariante und im zweiten Vergleich zwei Ausgestaltungsvarianten (ohne Bezug auf das Ausgangsmodell) miteinander verglichen,

4.3.1 Variante 1: Status-quo-Modell im Anhörungsdokument zur Festlegung

Ausgangsmodell der Betrachtung ist der Status quo des Klassifikationsmodell 2010 vor Aufnahme neuer ICD in das Modell, d. h. das Modell „Status quo“ der Erläuterung zum Entwurf der Festlegung (Anhörungsdokument[4] S. 49, Tabelle 9), vgl. Abbildung 4-5.

4.3.2 Variante 2: Anhörungsvorschlag (Modellvorschlag in den Festlegungen)

Variante 2 bildet den Anhörungsvorschlag ab, d. h. im Vergleich zum Status quo werden die DxG826 und DxG827 in die HMG202, die DxG825 und die DxG124 in die HMG021 sowie die DxG828 in die HMG022 eingruppiert. Variante 2 entspricht dem „Modellvorschlag“ in der Erläuterung zur Festlegung von Morbiditätsgruppen, Zuordnungsalgorithmus, Regressionsverfahren und Berechnungsverfahren, S. 60, Tab. 5), vgl. Abb. 4-6.

4.3.3 Variante 3: Endgültige Festlegung

Variante 3 bildet die endgültige Festlegung der Hierarchiestruktur ab. Im Vergleich zum Festlegungsentwurf wird die HMG202 um die Diagnosen E76.2 und E76.3 (mit ERT/SRT[5]) erweitert, der HMG021 wird die Diagnose E76.2 (ohne ERT/SRT) zugeord-

[4] Achtung: nur an dieser Stelle wird auf Entwurf der Festlegung vom 30.07.2010 Bezug genommen, die weiteren Verweise beziehen sich auf das eigentliche Festlegungsdokument vom 30.09.2010.

[5] ERT = Enzymersatztherapie, SRT = Substrat-Reduktionstherapie

net, die neue HMG225 wird zwischen den HMG202 und HMG021 einsortiert. Dies entspricht dem Vorschlag „Anhörung 2“ in der Erläuterung zur Festlegung von Morbiditätsgruppen, Zuordnungsalgorithmus, Regressionsverfahren und Berechnungsverfahren (S. 61, Tabelle 6), vgl. Abbildung 4-7.

Abbildung 4-5: Hierarchie „Metabolische Erkrankungen“ in der Status quo-Variante des Anhörungsdokuments

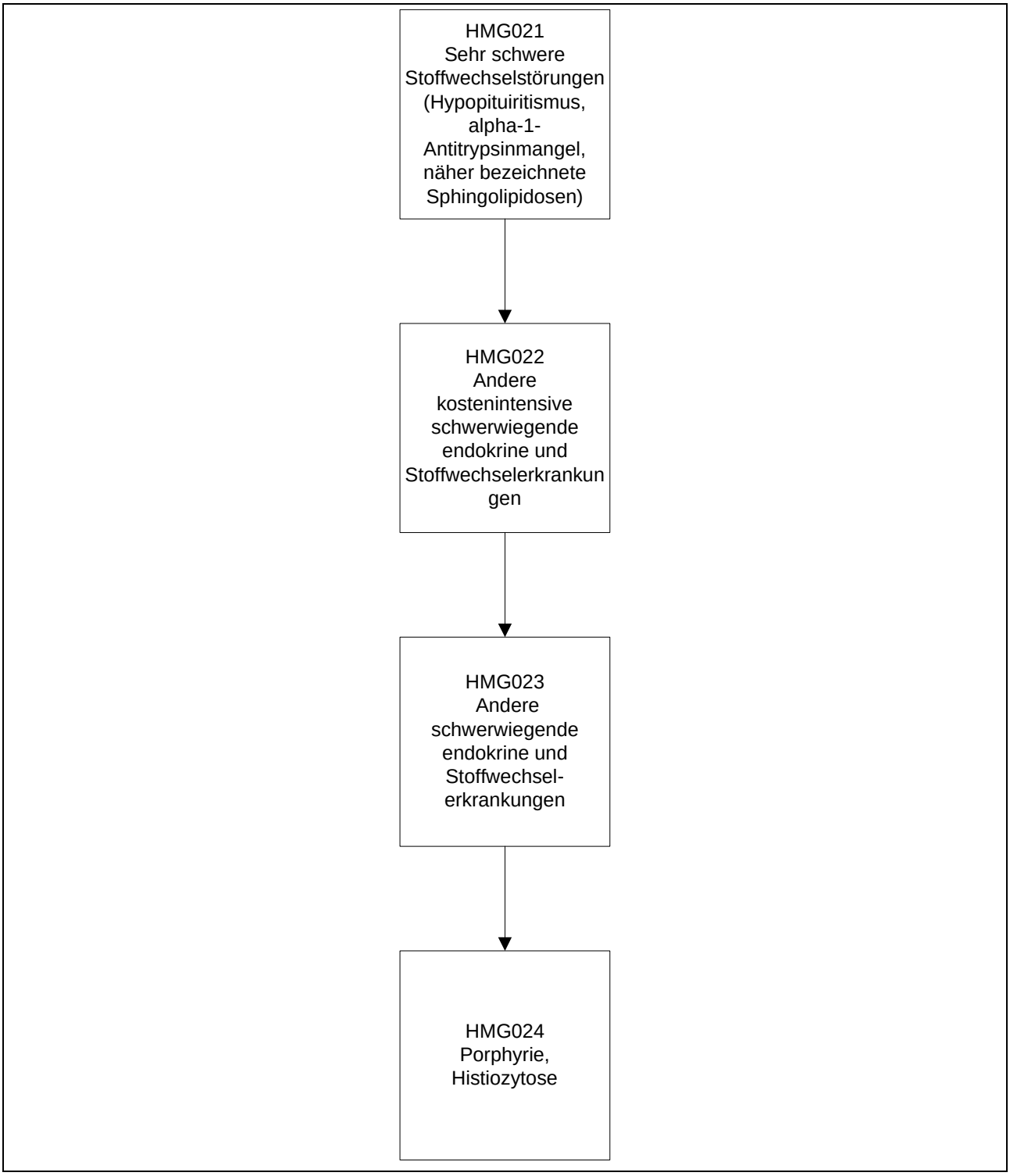


Abbildung 4-6: Hierarchie "Metabolische Erkrankungen" im Entwurf zur Festlegung

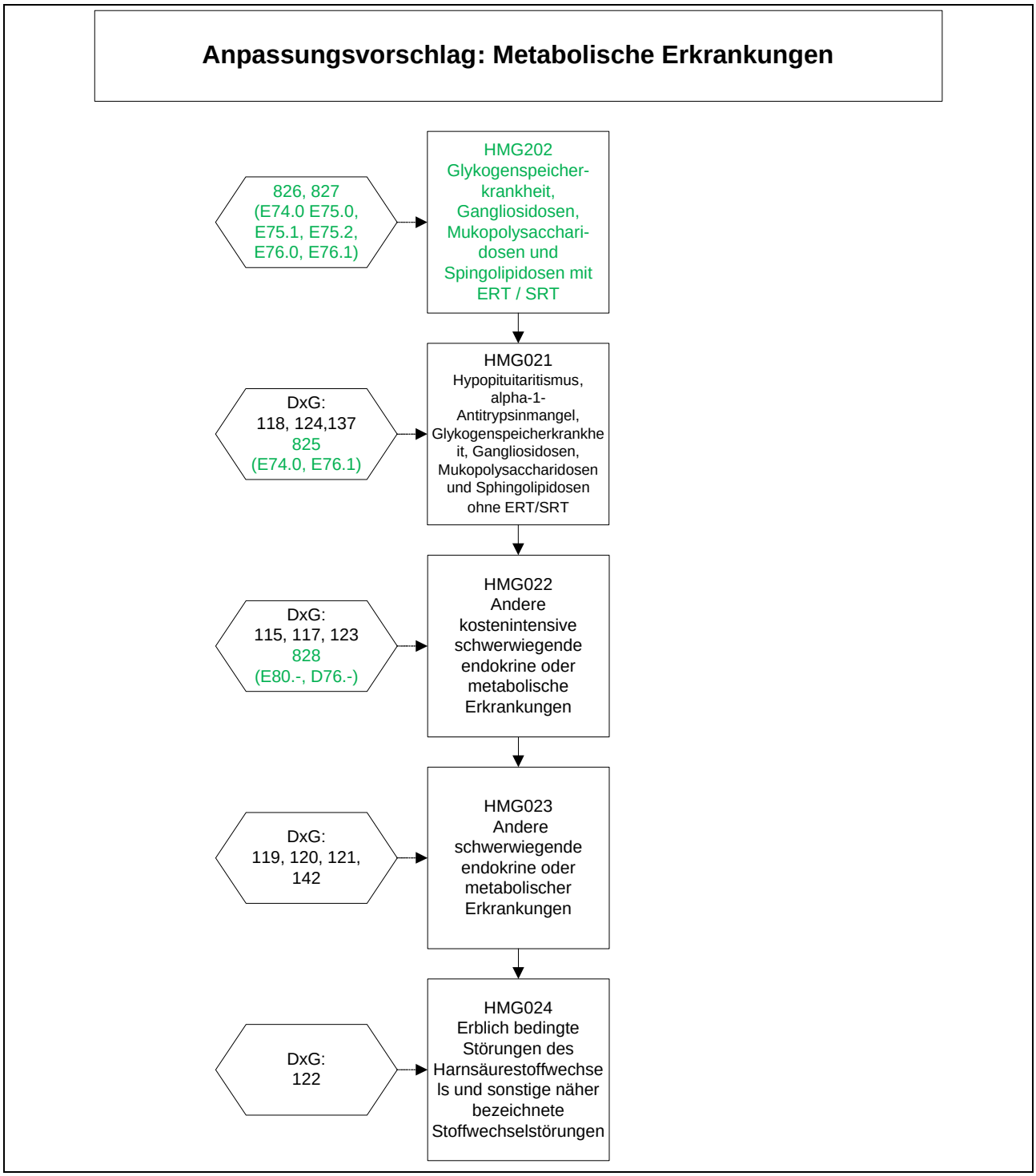
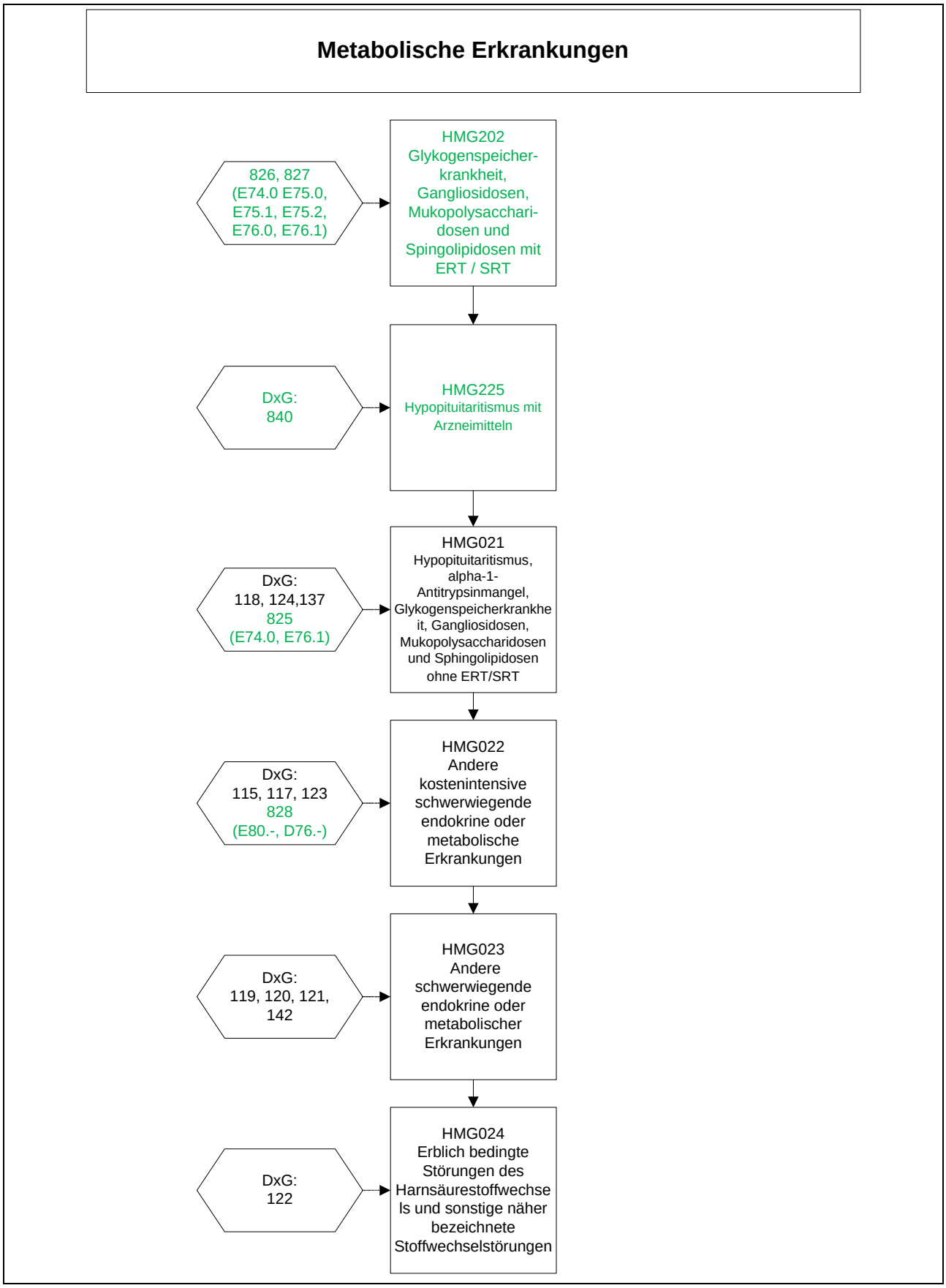


Abbildung 4-7: Hierarchie „Metabolische Erkrankungen“ in der endgültigen Festlegung



4.3.4 Definition von „betroffen“ und realisierte Stichprobenumfänge

In der weiten Definition werden alle diejenigen Versicherte als betroffen eingestuft („Betroffene I“), denen mindestens eine HMGs der Hierarchie „Metabolische Erkrankungen“ zugeordnet ist (HMG021 - HMG024, HMG202).

In der engeren Definition in Bezug auf die Variante 2 werden nur Versicherte als betroffen eingestuft („Betroffene II“), die von den im Festlegungsentwurf vorgeschlagenen Änderungen (Veränderung Variante 1 zu Variante 2) unmittelbar betroffen sind (Versicherte mit DXG124 oder DxG825 – DxG828).

In der engeren Definition in Bezug auf die Variante 3 („Betroffene III“) werden nur Versicherte als betroffen eingestuft, die von den in der endgültigen Festlegung umgesetzten Änderungen im Vergleich zum Festlegungsentwurf (Veränderung Variante 2 zu Variante 3) betroffen sind (Versicherte mit ICD E76.2, E76.3 mit ERT/SRT, E76.2 ohne ERT/SRT oder DxG840).

Die folgende Tabelle weist die verschiedenen Stichprobenumfänge aus.

Tabelle 4-3: Stichprobenumfänge der Bewertungsstichproben für die Vergleiche in der Hierarchie der metabolischen Erkrankungen*)

	q = 1			q = 2		q = 3	
Vergleich	N_B	N_{NB}	m	N_{NB}	m	N_{NB}	m
V 1 mit V 2, Betroffene I	37.625	37.625	72.250				
V 1 mit V 2, Betroffene II	7.710	7.710	15.420	15.420	23.130	23.130	30.840
V 2 mit V 3, Betroffene I	37.625	37.625	72.250				
V 2 mit V 3, Betroffene III	360	2.000	2.360	4.000	4.360	6.000	6.360

*) N_B = Zahl der Betroffenen, N_{NB} = Umfang der Stichprobe aus den Nichtbetroffenen
m = Umfang der Bewertungsstichprobe

Was die Zahl p der Prädiktoren anbetrifft, so gilt p=152 für die erste, p=153 für die zweite und p=154 für die dritte Variante..

5 Ergebnisse der Erprobung

5.1 Optimale Wahl von N_{NB} bei enger Operationalisierung von „betroffen“

Insgesamt wurde die Ziehung und Zusammenstellung der Bewertungsstichprobe für jeden Vergleich 9.000 mal wiederholt. Die erste Untersuchung galt der Frage, wie groß N_{NB} im Verhältnis zu N_B gewählt werden sollte (d. h. mit welchem q am besten gearbeitet wird), wenn eine enge Operationalisierung des Begriffs „betroffen“ verwendet wird. Für diese Untersuchung wurden alle 9.000 Wiederholungen verwendet (n=9.000) und die Mittelwerte der Differenzen von R^2 und CPM ausgewertet (sowie ggf. BIC), d. h. es wurde anstelle von MAPE die im Allgemeinen mit wesentlich kleineren Differenzen einhergehende Maßzahl CPM verwendet.

5.1.1 Erkrankungen der Lunge

In Tabelle 5-1 wird der Mikroskopeffekt ausgewiesen. Dieser ist berechnet als das Verhältnis des Mittelwertes der jeweiligen Differenz aus den n=9.000 Bewertungsstichproben zu der analogen Differenz, berechnet aus dem vollen Datensatz.

Tabelle 5-1: Mikroskopeffekte bei den Vergleichen von Varianten der Hierarchie „Erkrankungen der Lunge“ bei enger Operationalisierung von „betroffen“, Mittelwerte, n=9.000

Vergleich	Datensatz*)	$R^2_2 - R^2_1$	$CPM_2 - CPM_1$
	Voller Datensatz	-0,0000032	-0,0000027
V 2 mit V 1	MD, q = 1	-0,0010198	0,0091847
	MD, q = 2	-0,0009002	0,0072318
	MD, q = 3	-0,0008099	0,0060161
		Mikroskopeffekt	Mikroskopeffekt
V 2 mit V 1	MD, q = 1	319	-3.402
	MD, q = 2	281	-2.678
	MD, q = 3	253	-2.228
Vergleich	Datensatz	$R^2_3 - R^2_1$	$CPM_3 - CPM_1$
	Voller Datensatz	0,0002340	0,0001322
V 3 mit V 1	MD, q = 1	0,0708751	0,0769409
	MD, q = 2	0,0617855	0,0603831
	MD, q = 3	0,0556380	0,0502495
		Mikroskopeffekt	Mikroskopeffekt
V 3 mit V 1	MD, q = 1	303	582
	MD, q = 2	264	457
	MD, q = 3	238	380

*) MD = Mikroskopdesign

Aus der Tabelle geht zunächst hervor, dass beide Maßzahlen, wenn man sie auf dem vollen Datensatz berechnet, zur gleichen Entscheidung führen. Es erweist sich nämlich auf der Basis jeder der beiden Differenzen die Variante 1 besser als die Variante 2 und die Variante 3 besser als die Variante 1.

Es fällt sodann auf, dass die Bewertungen auf Basis der mittleren Differenzen, berechnet aus den Bewertungsstichproben des Mikroskop-Designs nur bezogen auf die Maß-

zahl R^2 zum gleichen Ergebnis kommen. Auf der Basis von CPM führt die Bewertungsmethodik des Mikroskop-Designs für den Vergleich der Variante 2 mit der Variante 1 zu dem entgegengesetzten Ergebnis: Wir finden für das Verhältnis ein negatives Vorzeichen, d. h. die Differenzen haben in den Bewertungsstichproben ein positives Vorzeichen, Variante 2 wird gegenüber Variante 1 (dem Ausgangsmodell) bevorzugt. Diese Problematik unterschiedlicher Ergebnisse, je nachdem, auf welche Maßzahl man sich bei der Bewertung stützt, wird in Abschnitt 5.6 näher erörtert.

Darüber hinaus zeigt sich, dass der Mikroskopeffekt für R^2 und beide Vergleiche am stärksten bei der Wahl von q=1 ausfällt. Hinsichtlich CPM gilt das auch für den Vergleich der Varianten 3 und 1. Was den Vergleich der Varianten 2 und 1 auf der Basis von CPM betrifft, kann man – angesichts des Vorzeichenwechsels der Differenz beim Übergang vom vollen Datensatz zur Bewertungsstichprobe – von Mikroskopeffekt nur sprechen, wenn man auf den absoluten Betrag der Differenzen abhebt. In diesem Sinn verstanden gilt aber auch für CPM, dass q=1 die beste Wahl ist. Ferner ist der Mikroskopeffekt in diesem Sinn wesentlich stärker für CPM als für R^2.

5.1.2 Neubildungen

Tabelle 5-2: Mikroskopeffekte bei den Vergleichen von Varianten der Hierarchie „Neubildungen“ bei enger Operationalisierung von „betroffen“, Mittelwerte, n=9.000

Vergleich	Datensatz*)	$R^2_2 - R^2_1$	$CPM_2 - CPM_1$
	Voller Datensatz	0,0000050	0,0000110
V 2 mit V 1	MD, q = 1	0,0026885	-0,0724092
	MD, q = 2	0,0023401	-0,0556384
	MD, q = 3	0,0020765	-0,0456185
		Mikroskopeffekt	Mikroskopeffekt
V 2 mit V 1	MD, q = 1	538	-6.583
	MD, q = 2	468	-5.058
	MD, q = 3	415	-4.147
Vergleich	Datensatz	$R^2_3 - R^2_2$	$CPM_3 - CPM_2$
	Voller Datensatz	0,0000366	0,0000165
V 3 mit V 2	MD, q = 1	0,0815755	0,0421232
	MD, q = 2	0,0406985	0,0220749
	MD, q = 3	0,0267872	0,0149561
		Mikroskopeffekt	Mikroskopeffekt
V 3 mit V 2	MD, q = 1	2.229	2.553
	MD, q = 2	1.112	1.338
	MD, q = 3	732	906

*) MD = Mikroskopdesign

Auch für die Hierarchie der Neubildungen gilt, dass beide Maßzahlen, wenn man sie auf dem vollen Datensatz berechnet, zu den gleichen Entscheidungen führen, so dass Variante 2 gegenüber Variante 1 und Variante 3 gegenüber Variante 2 bevorzugt werden.

Der Vergleich der Varianten 2 und 1 auf der Basis von CPM ist wie in der Hierarchie „Erkrankungen der Lunge“ durch einen Vorzeichenwechsel beim Übergang vom vollen

Datensatz zu den Bewertungsstichproben gekennzeichnet. Die Erörterungen im vorigen Abschnitt gelten hier entsprechend. Ferner zeigt Tabelle 5-2 (wie schon Tabelle 5-1), dass die Verhältniszahlen für beide betrachtete Maßzahlen und beide Vergleiche am größten bei Wahl von q=1 ausfällt (und, was die Maßzahlen betrifft, größer für CPM als für R^2).

5.1.3 Metabolische Erkrankungen

Tabelle 5-3: Mikroskopeffekte bei den Vergleichen von Varianten der Hierarchie „Metabolische Erkrankungen“ bei enger Operationalisierung von „betroffen“, Mittelwerte, n=9.000

Vergleich	Datensatz*)	$R^2_2 - R^2_1$	$CPM_2 - CPM_1$	$BIC_2 - BIC_1$
	Voller Datensatz	0,01757	0,00094	-0,02316
V 2 mit V 1	MD, q = 1	0,39685	0,04469	-0,61400
	MD, q = 2	0,37984	0,03759	-0,58734
	MD, q = 3	0,36555	0,03263	-0,56308
		Mikroskopeffekt	Mikroskopeffekt	Mikroskopeffekt
V 2 mit V 1	MD, q = 1	23	48	27
	MD, q = 2	22	40	25
	MD, q = 3	21	35	24
Vergleich	Datensatz	$R^2_3 - R^2_2$	$CPM_3 - CPM_2$	$BIC_3 - BIC_2$
	Voller Datensatz	0,00373	0,00043	-0,00498
V 3 mit V 2	MD, q = 1	0,53570	0,08982	-0,82269
	MD, q = 2	0,50032	0,06781	-0,76673
	MD, q = 3	0,47031	0,05515	-0,71512
		Mikroskopeffekt	Mikroskopeffekt	Mikroskopeffekt
V 3 mit V 2	MD, q = 1	144	209	165
	MD, q = 2	134	158	154
	MD, q = 3	126	128	144

*) MD = Mikroskopdesign

In dieser Tabelle stimmen sowohl im oberen Teil (Vergleich von V 2 mit V 1) als auch im unteren Teil (Vergleich von V 3 mit V 2) alle Bewertungen überein, unbeschadet der Tatsache, auf welche Maßzahl sie gestützt werden und ob sie auf dem vollen Datensatz oder auf den 9.000 Bewertungsstichproben basieren.

Der Mikroskopeffekt ist im Vergleich zwischen den drei Hierarchien der kleinste, aber wiederum am größten für q=1 (und, was die Maßzahlen betrifft, am größten für CPM).

5.1.4 Zwischenresümee

Die Überhöhung des Stichprobenumfangs der Stichprobe der Nichtbetroffenen gegenüber der Zahl der Betroffenen ist auch bei sehr kleinen Zahlen der Betroffenen nicht erforderlich, man kann beim balancierten Design (mit der Nebenbedingung $N_{NB} \geq 2.000$) bleiben. Denn die Ergebnisse ändern sich nicht durch Erhöhung des q und der Mikroskopeffekt wird dadurch kleiner. Im Folgenden werden daher nur noch die Auswertungen für den Fall q = 1 präsentiert.

5.2 Verteilungen der Kennziffern und der Differenzen

Da geplant ist, die Untersuchungen der Differenzen auf die Mittelwerte der Verteilungen zu stützen, ist zunächst zu klären, ob der Mittelwert als ein geeigneter Repräsentant der jeweiligen Verteilung angesehen werden kann. Eine weitere interessante Frage richtet sich auf die Unterschiede, die zwischen den Verteilungen einer bestimmten Kennziffer oder Differenz bestehen, wenn man das Betroffenen-Konzept variiert. Es zeigt sich, dass die Unterschiede der Verteilungen der Kennziffern bei gleicher Operationalisierung von „betroffen“ zwischen den Varianten der gleichen Hierarchie minimal sind.

5.3 Histogramme

5.3.1 Maßzahlen

Die im Folgenden präsentierten und erörterten Histogramme werden auf der Basis der 9.000 Bewertungsstichproben dargestellt. Wir beschränken uns dabei, was die Verteilungen der Kennziffern betrifft, auf jeweils eine Variante. Aber selbst dann handelt es sich noch um 2x3x2 = 12 Histogramme (zwei Kennziffern, drei Hierarchien und zwei Betroffenen-Konzepte). Dazu kommen noch 6x2 = 12 Differenzen. Allerdings gibt es viele Ähnlichkeiten unter diesen 24 Histogrammen, so dass es genügt, einige typische Beispiele zu erörtern.

Der Wechsel des Betroffenen-Konzepts wirkt sich in erheblichen Umfang auf die Skalen des jeweiligen Histogramms aus. Um dennoch einen sachgerechten Vergleich zu ermöglichen, wurden zwei Histogramme, welche sich nur in der Wahl des Betroffenen-Konzepts unterscheiden, mit einer gemeinsamen Skala auf der y-Achse dargestellt (dort sind absolute Häufigkeiten, also Versichertenzahlen ausgewiesen). In einer zweiten Version wurde dann auch eine gemeinsame x-Achse festgelegt (dort sind R^2/CPM-Werte bzw. deren Differenzen ausgewiesen).

Betrachten wir zunächst die Verteilung von R^2 am Beispiel der Variante 2 der Hierarchie „Neubildungen“ (vgl. Abbildung 5-1) so zeigt sich eine geringfügig schiefe Verteilung. Außerdem liegt ihre Wölbung, insbesondere bei Zugrundelegung der engen Operationalisierung, deutlich über der eingezeichneten „passenden“ Normalverteilung mit gleichem Mittelwert und gleicher Standardabweichung.[6] Wird das Mikroskop-Design mit dem engen Betroffenen-Konzept realisiert statt mit dem weiten, so ändert sich die Lage wenig, aber die Standardabweichung wird erheblich größer (vgl. die unteren beiden Diagramme von Abbildung 5-1).[7] Gleichzeitig sind die mittleren Histogrammklassen um den Modalwert herum viel stärker besetzt und die Besetzungszahlen fallen zu den Randklassen

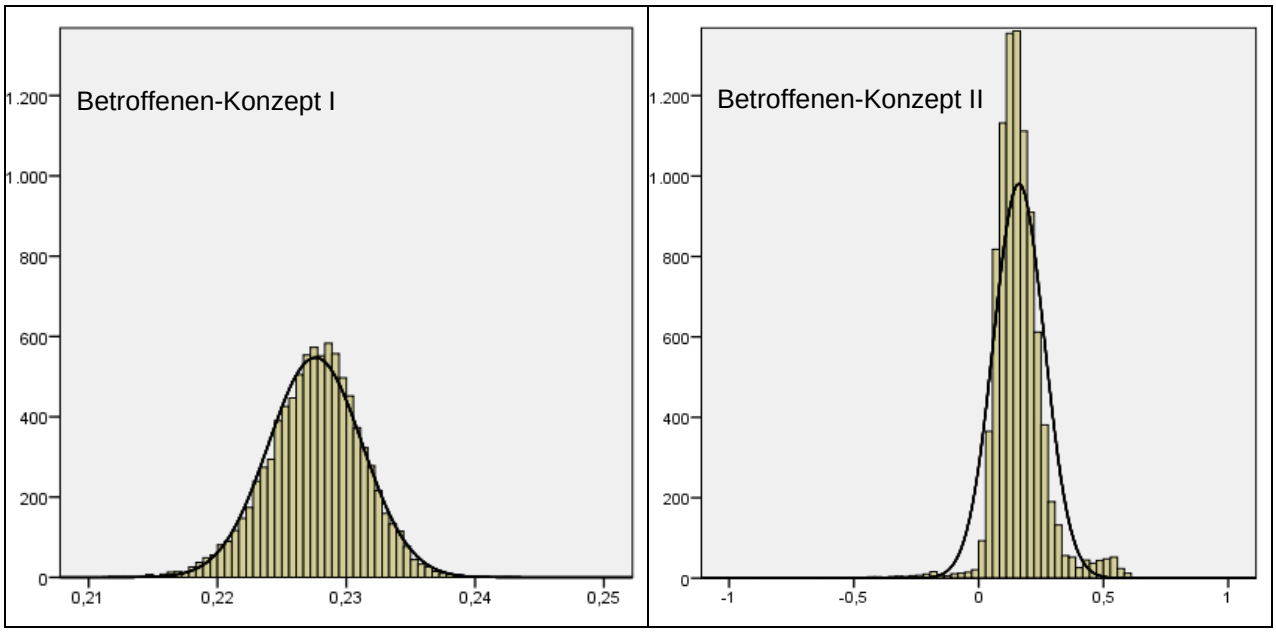
[6] Die Schiefe einer Verteilung ist definiert als ihr auf die dritte Potenz der Standardabweichung normiertes drittes zentrales Moment. Die Wölbung (Kurtosis) einer Verteilung ist definiert als ihr auf die vierte Potenz der Standardabweichung normiertes viertes zentrales Moment. Jede Normalverteilung hat die Schiefe Null und die Wölbung 3.

[7] Der Faktor, um den sie größer wird, beträgt im diskutierten Beispiel 27,9. Die Vergrößerung liegt natürlich an dem um den Faktor 1/94,3 kleineren Stichprobenumfang jeder der Bewertungsstichproben (vgl. Tabelle 4-2).

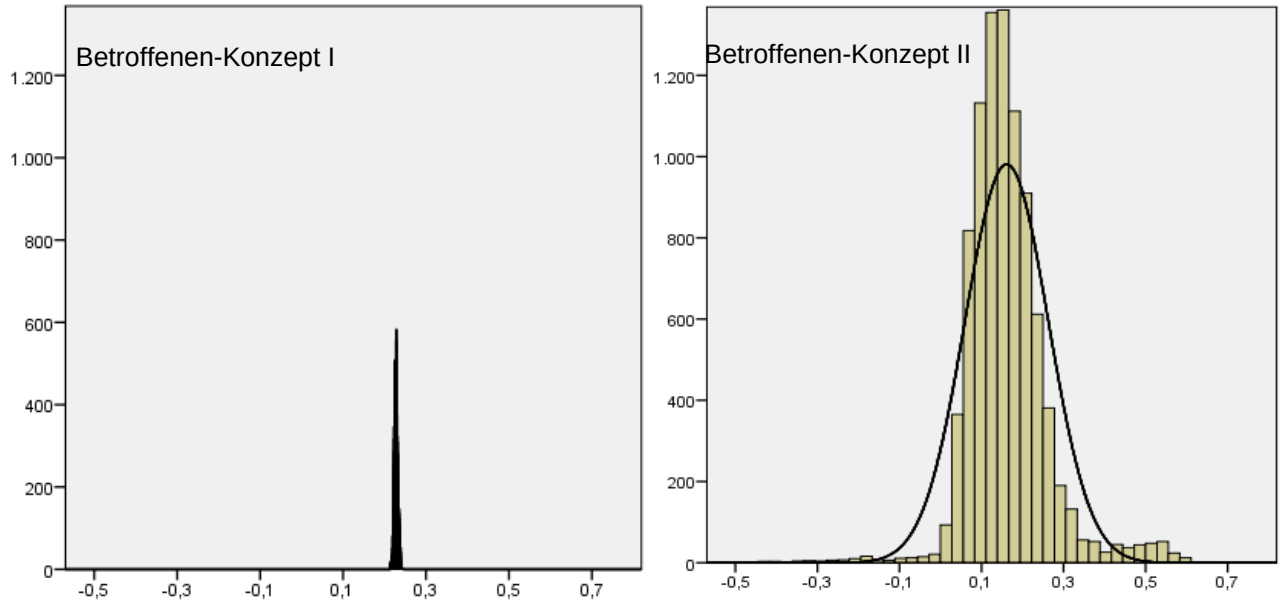
sehr viel stärker ab. Die größere Standardabweichung führt auch dazu, dass (bei positivem Mittelwert) die R^2-Werte für eine kleine Teilmenge der Bewertungsstichproben negativ ausfallen (im Beispiel für 140 der 9.000 Bewertungsstichproben). Dies ist nur möglich, weil das Modell am vollen Datensatz kalibriert, R^2 aber jeweils an der kleinen Bewertungsstichprobe berechnet wurde.

Abbildung 5-1: Histogramme der Verteilung von R^2 der Variante 2 der Hierarchie „Neubildungen“ für die weite (linkes Bild) und die enge Definition der Betroffenen (rechtes Bild)

Individuelle Skalierungen auf der x-Achse



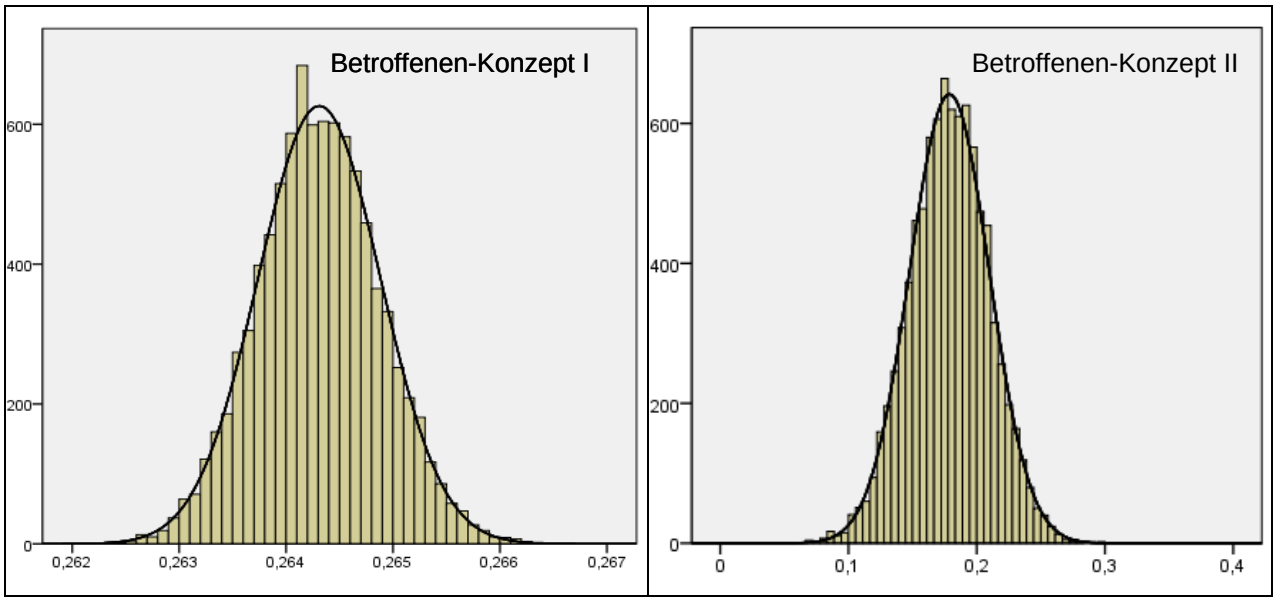
Gleiche Skalierungen auf der x-Achse



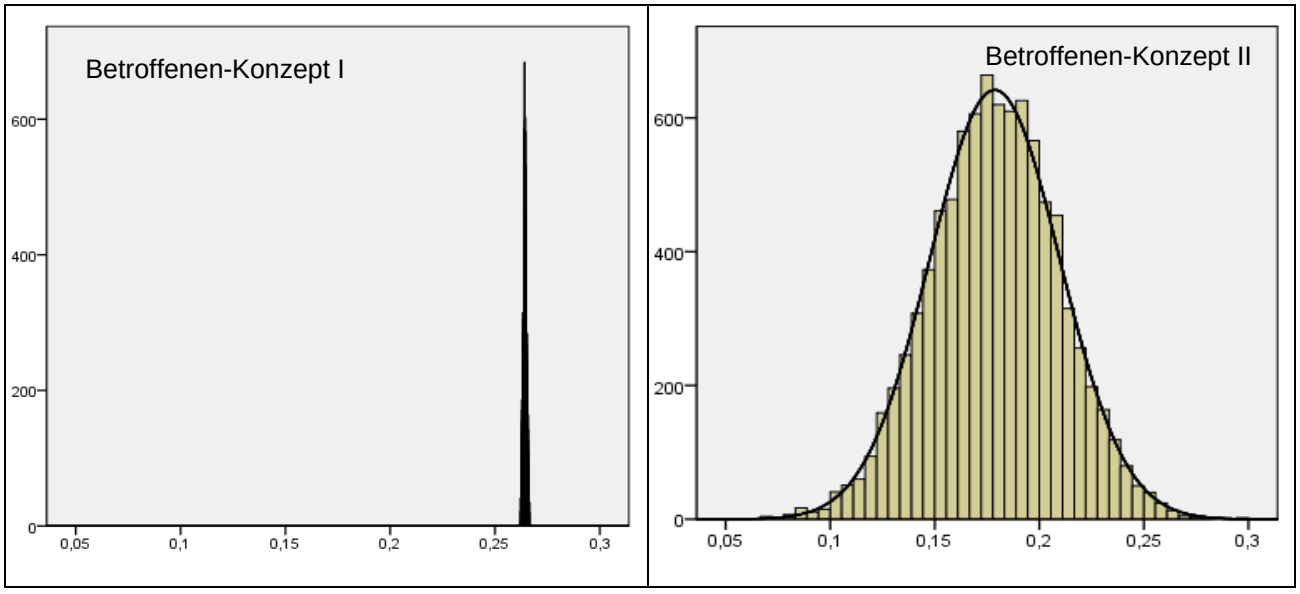
CMP zeigt ein anderes Bild. Für beide Betroffenen-Konzepte sind die mittleren Klassen gleich stark besetzt und der Abfall zu den Randklassen vollzieht sich wie in der passenden Normalverteilung (vgl. Abbildung 5-2). Allerdings ist die Standardabweichung bei enger Operationalisierung von „betroffen“ (bei viel niedrigeren Ausgangsniveau) wegen des um den Faktor 1/94,3 kleineren Stichprobenumfangs jeder der Bewertungsstichproben wiederum um einen erheblichen Faktor größer, als bei der weiten.

Abbildung 5-2: Histogramm der Verteilung von CPM der Variante 2 der Hierarchie „Neubildungen“ für die weite (links) und die enge Definition der Betroffenen (rechtes Bild)

Individuelle Skalierungen auf der x-Achse



Gleiche Skalierungen auf der x-Achse



5.3.2 Differenzen

Wichtiger (für die Bewertung) als die Verteilungen der Kennziffern selbst sind die Verteilungen der Differenzen. Diese untersuchen wir zunächst am Beispiel des Vergleichs zwischen den Varianten 2 und 1 in der Hierarchie Neubildungen, weil sich auf der Basis von CPM (nicht aber auf derjenigen von R^2) beim Übergang vom weiten zum engen Betroffenen-Konzept die Bewertung umdreht (vgl. Abschnitt 5.1.2).

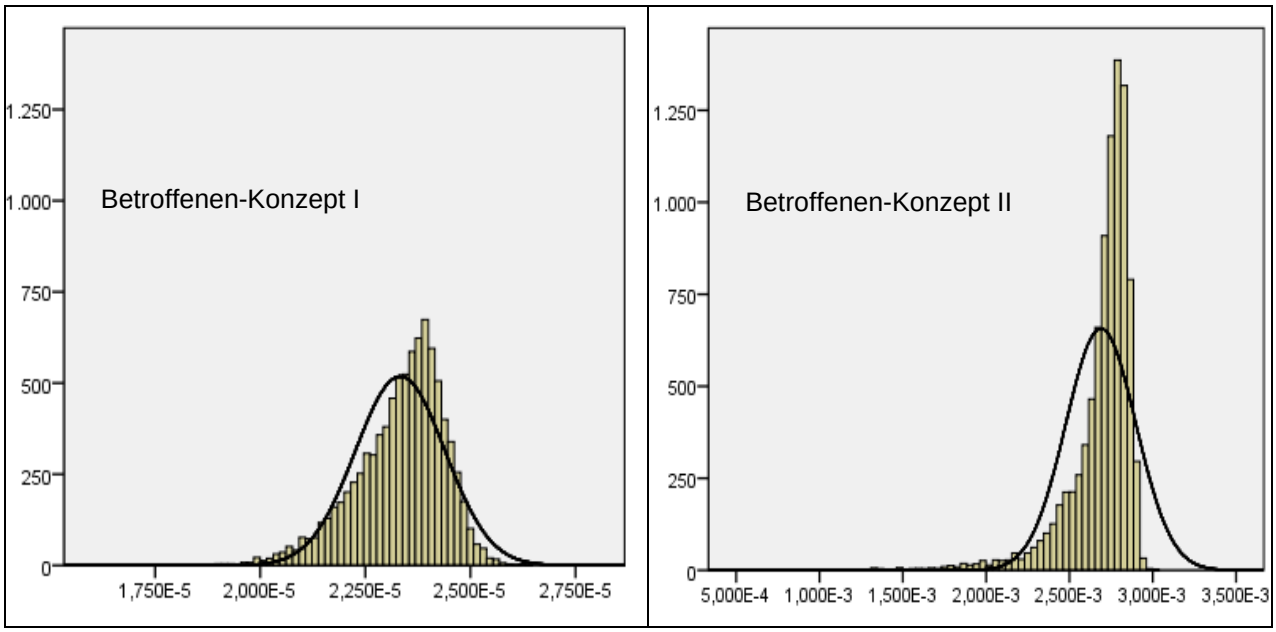
Die Verteilung der Differenz der R^2-Werte ist links schief (rechts steil), während diejenige der CPM-Werte (insbesondere für das Betroffenen-Konzept I) annährend symmetrisch ist. Die Unterschiede im Klassenbesetzungsmuster (mittlere gegen Randklassen) ähneln den für die Kennziffern beschriebenen (vgl. Abbildung 5-3 und Abbildung 5-4)

Während die Verteilungen der Differenzen sowohl von den R^2-Werten als auch von den CPM-Werten bei Realisierung des weiten Betroffenen-Konzepts I etwa im Bereich 10^{-5} angesiedelt sind, finden wir bei Realisierung des engen Betroffenen-Konzepts große Unterschiede: Die Verteilung der Differenz von R^2 liegt nun im Bereich von 10^{-3}, also weiter von Null weg, während die Verteilung der Differenz von CPM über die Null hinaus in den negativen Bereich geschoben wird. In beiden Fällen geht die Veränderungen mit einer erheblichen Vergrößerung der Standardabweichung einher. Diese ist jedoch im Hinblick auf die Lage des Mittelwertes zu relativieren. Bezieht man die Standardabweichung auf den Mittelwert, so erhält man den Variationskoeffizient der Verteilung, der üblicherweise in Prozent angegeben wird. Die Verteilung der Differenz von CPM besitzt nun für beide Betroffenen-Konzepte einen deutlich kleineren Variationskoeffizient und dieser wächst bei Änderung des Betroffenen-Konzepts von I auf II auch nicht so stark an, wie Tabelle 5-6 (S. 43) ausweist. Auf diese Beobachtungen werden wir in Abschnitt 5.4 noch näher eingehen.

Da für Vergleiche im Bereich der Hierarchie „Metabolische Erkrankungen“ auch die Differenz der BIC-Werte herangezogen wird, ist die Verteilung dieser Differenz ebenfalls von Interesse. Diese ähnelt der Verteilung der Differenz von R^2, nur das sie nicht links-, sondern rechtsschief ist (vgl. Abbildung 5-5).

Abbildung 5-3: Histogramm der Verteilung der Differenzen der R^2-Werte zwischen den Varianten 2 und 1 der Hierarchie „Neubildungen“ für die weite (links) und die enge Definition der Betroffenen (rechtes Bild)

Individuelle Skalierungen auf der x-Achse



Gleiche Skalierungen auf der x-Achse

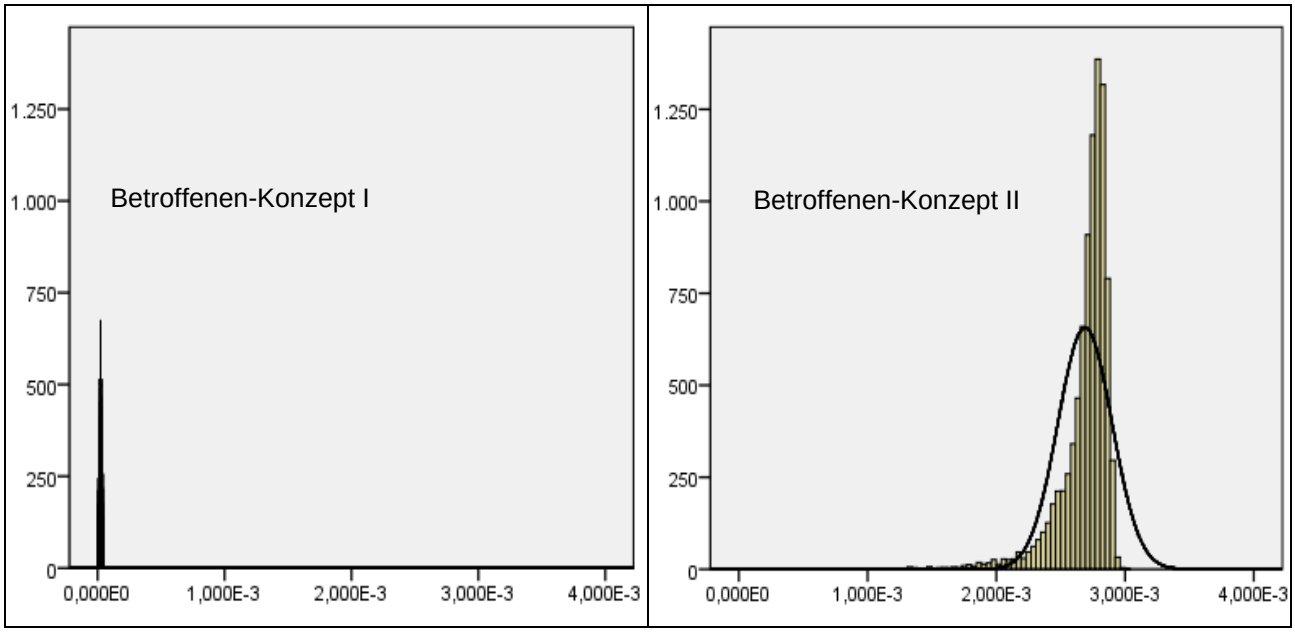
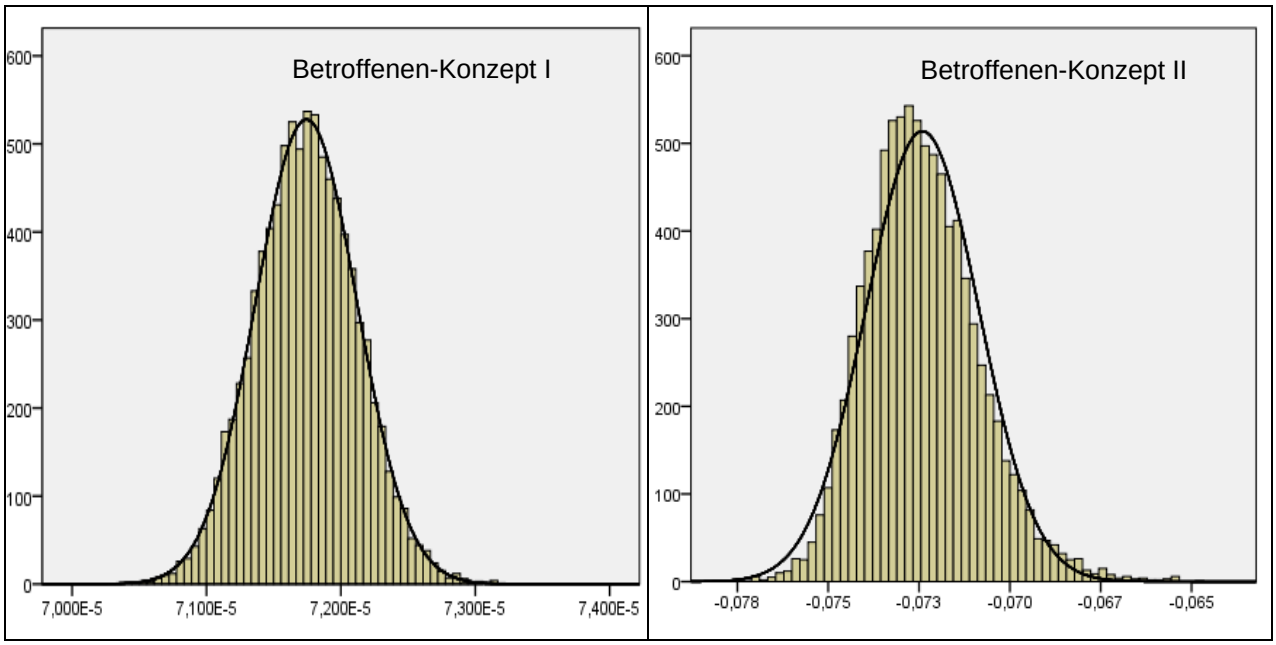


Abbildung 5-4: Histogramm der Verteilung der Differenzen der CPM-Werte zwischen den Varianten 2 und 1 der Hierarchie „Neubildungen“ für die weite (linkes Bild) und die enge Definition der Betroffenen (rechtes Bild)

Individuelle Skalierungen auf der x-Achse



Gleiche Skalierungen auf der x-Achse

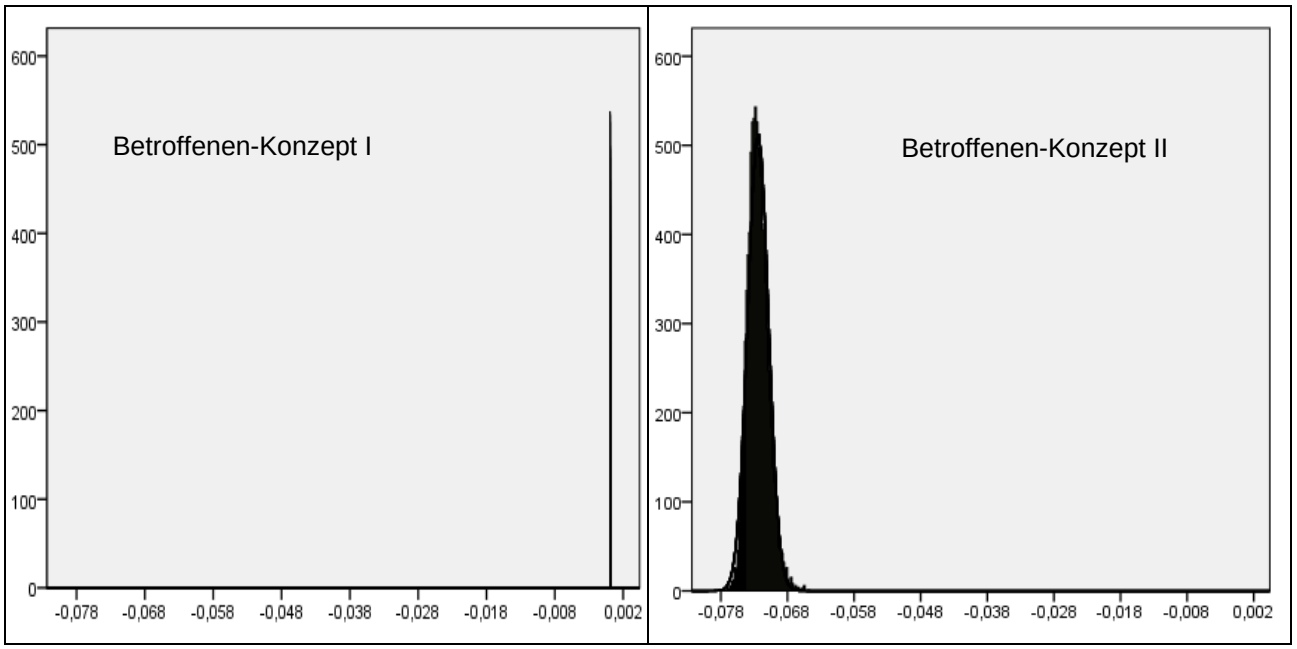
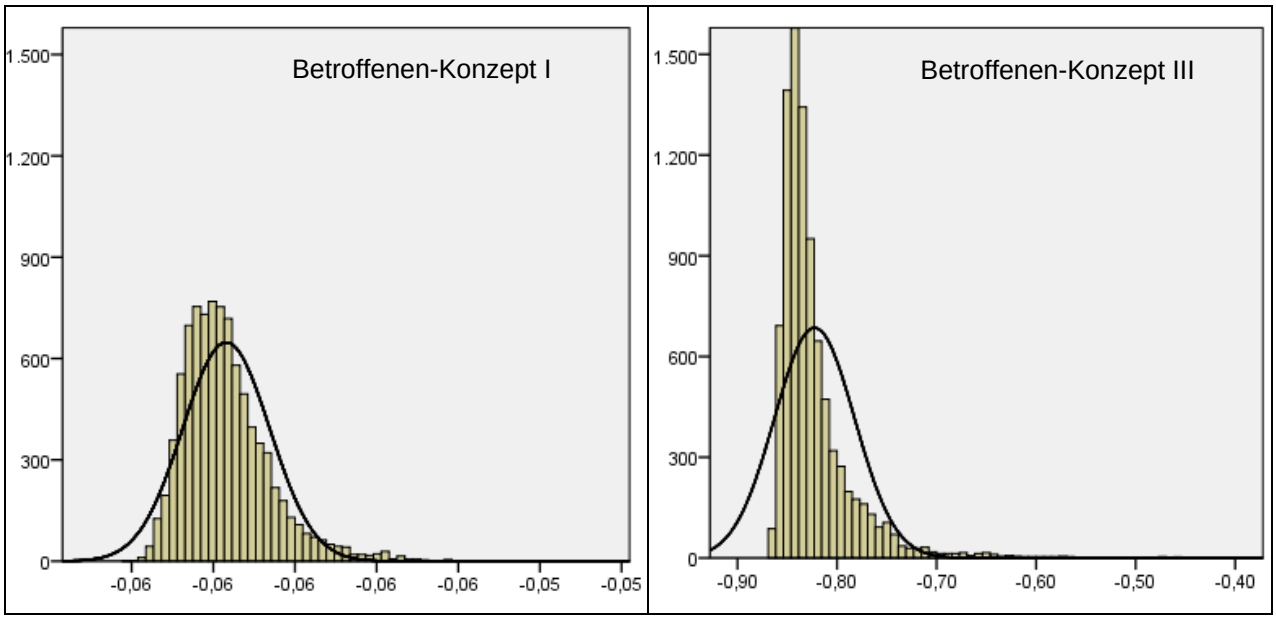
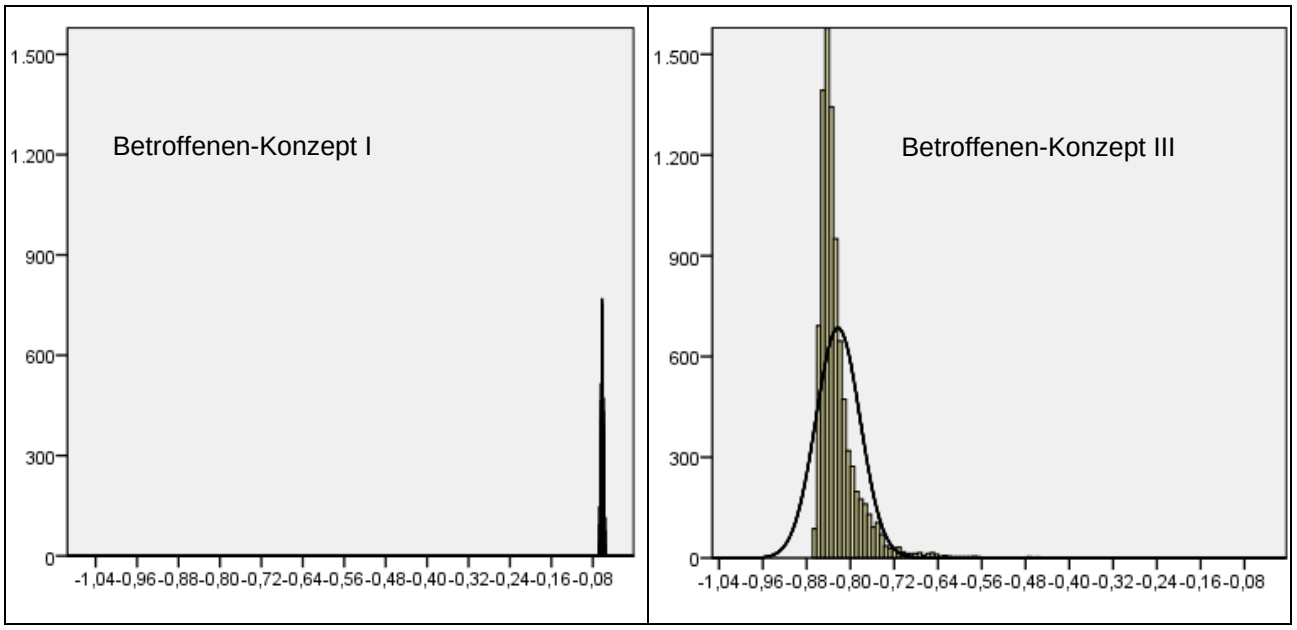


Abbildung 5-5: Histogramm der Verteilung der Differenzen der BIC-Werte zwischen den Varianten 3 und 2 der Hierarchie „Metabolische Erkrankungen“ für die weite (linkes Bild) und die enge Definition der Betroffenen (rechtes Bild)

Individuelle Skalierungen auf der x-Achse



Gleiche Skalierungen auf der x-Achse



5.4 Verteilungsparameter der Differenzen

5.4.1 Erläuterungen und Resümee

In diesem Abschnitt werden für die sechs untersuchten Vergleiche folgende Verteilungsparameter der Differenzen der R^2- und CPM-, sowie ggf. der BIC-Werte zusammengestellt:

- Median, Mittelwert, Standardabweichung, Variationskoeffizient, Schiefe und Standardfehler der Schiefe

Mit Hilfe des Standardfehlers der Schiefe lässt sich beurteilen, ob die ausgewiesene Schiefe statistisch signifikant von Null verschieden ist. Auf dem 5%-Niveau gilt dabei die Faustregel, dass Signifikanz vorliegt, wenn der Absolutbetrag der Schiefe mindestens doppelt so groß ist wie ihr Standardfehler.

Analog lässt sich die Signifikanz des Mittelwertes überprüfen, wobei sich der Standardfehler des Mittelwertes ergibt, indem man die ausgewiesene Standardabweichung durch $\sqrt{9.000}$ dividiert.

Da die Standardabweichungen klein und der Stichprobenumfang groß ist, erweisen sich fast alle Lage- und Schiefeparameter als signifikant. Nicht signifikante Ergebnisse stellen die Minderheit dar, betreffen ausschließlich die Schiefe der Verteilung von CPM und wurden in den Tabellen grau unterlegt.

Im Resümee kann festgehalten werden, dass trotz einer gewissen Schiefe der Verteilungen, insbesondere der R^2- und der BIC-Differenzen, Median und Mittelwert stets so nah beieinander liegen, dass man den Mittelwert als Repräsentant der Lage der Verteilung ansehen kann.

Darüber hinaus fällt auf, dass die Variationskoeffizienten insgesamt sehr klein sind und in allen Fällen für die CPM-Differenzen deutlich kleiner ausfallen, als für die R^2-Differenzen.

5.4.2 Erkrankungen der Lunge

Tabelle 5-4: Maßzahlen der Verteilung der Differenzen von R^2 und CPM beim Vergleich der Varianten 2 und 1 in der Hierarchie „Erkrankungen der Lunge“

a) Betroffenen-Konzept I

Verteilung der Differenz	Median	Mittelwert	Standard-abwei-chung	Variations-koeffizient	Schiefe	Standard-fehler der Schiefe
$R^2_2 - R^2_1$	-,000010	-,000010	,00000019	-1,8%	,295	,026
$CPM_2 - CPM_1$	-,000119	-,000119	,00000024	-0,2%	,017	,026

b) Betroffenen-Konzept II

Verteilung der Differenz	Median	Mittelwert	Standard-abwei-chung	Variations-koeffizient	Schiefe	Standard-fehler der Schiefe
$R^2_2 - R^2_1$	-,00104	-,00102	,000072	-7,1%	3,4	,026
$CPM_2 - CPM_1$	,00920	,00918	,000177	1,9%	-,541	,026

Tabelle 5-5: Maßzahlen der Verteilung der Differenzen von R^2 und CPM beim Vergleich der Varianten 3 und 1 in der Hierarchie „Erkrankungen der Lunge“

a) Betroffenen-Konzept I

Verteilung der Differenz	Median	Mittelwert	Standard-abwei-chung	Variations-koeffizient	Schiefe	Standard-fehler der Schiefe
$R^2_3 - R^2_1$	,00103	,00102	,0000192	1,9%	-,104	,026
$CPM_3 - CPM_1$	,00083	,00083	,0000066	0,8%	-,934	,026

b) Betroffenen-Konzept III

Verteilung der Differenz	Median	Mittelwert	Standard-abwei-chung	Variations-koeffizient	Schiefe	Standard-fehler der Schiefe
$R^2_3 - R^2_1$	0,072	0,071	,0048	6,8%	-3,2	,026
$CPM_3 - CPM_1$	0,077	0,077	,0016	2,0%	-,516	,026

5.4.3 Neubildungen

Tabelle 5-6: Maßzahlen der Verteilung der Differenzen von R^2 und CPM im Rahmen des Vergleichs der Varianten 2 und 1 in der Hierarchie „Neubildungen“

a) Betroffenen-Konzept I

Verteilung der Differenz	Median	Mittelwert	Standard-abwei-chung	Variations-koeffizient	Schiefe	Standard-fehler der Schiefe
$R^2_2 - R^2_1$	,000024	,000023	,00000107	4,6%	-,821	,026
$CPM_2 - CPM_1$	,000072	,000072	,00000038	0,5%	,050	,026

b) Betroffenen-Konzept II

Verteilung der Differenz	Median	Mittelwert	Standard-abwei-chung	Variations-koeffizient	Schiefe	Standard-fehler der Schiefe
$R^2_2 - R^2_1$	,0027	,0027	,00021	7,8%	-2,9	,026
$CPM_2 - CPM_1$	-,0725	-,0724	,00155	-2,1%	,525	,026

Tabelle 5-7: Maßzahlen der Verteilung der Differenzen von R^2 und CPM im Rahmen des Vergleichs der Varianten 3 und 2 in der Hierarchie „Neubildungen“

a) Betroffenen-Konzept I

Verteilung der Differenz	Median	Mittelwert	Standard-abwei-chung	Variations-koeffizient	Schiefe	Standard-fehler der Schiefe
$R^2_3 - R^2_2$	,000177	,000176	,00000270	1,5%	-,730	,026
$CPM_3 - CPM_2$	,000093	,000093	,00000022	0,2%	,001	,026

b) Betroffenen-Konzept III

Verteilung der Differenz	Median	Mittelwert	Standard-abwei-chung	Variations-koeffizient	Schiefe	Standard-fehler der Schiefe
$R^2_3 - R^2_2$	,083	,082	,0277	34,0%	-,10	,026
$CPM_3 - CPM_2$	,042	,042	,0033	7,9%	,014	,026

5.4.4 Metabolische Erkrankungen

Tabelle 5-8: Maßzahlen der Verteilung der Differenzen von R^2, CPM und BIC im Rahmen des Vergleichs der Varianten 2 und 1 in der Hierarchie „Metabolische Erkrankungen“

a) Betroffenen-Konzept I

Verteilung der Differenz	Median	Mittelwert	Standard-abweichung	Variations-koeffizient	Schiefe	Standard-fehler der Schiefe
$R^2_2 - R^2_1$	,1564	,1561	,001713	1,1%	-1,60	,026
$CPM_2 - CPM_1$	,0201	,0201	,000067	0,3%	-,093	,026
$BIC_2 - BIC_1$	-,2496	-,2489	,003697	-1,5%	1,32	,026

b) Betroffenen-Konzept II

Verteilung der Differenz	Median	Mittelwert	Standard-abweichung	Variations-koeffizient	Schiefe	Standard-fehler der Schiefe
$R^2_2 - R^2_1$	,398	,397	,005154	1,3%	-3,0	,026
$CPM_2 - CPM_1$	,045	,045	,000332	0,7%	-,28	,026
$BIC_2 - BIC_1$	-,6165	-,6140	,010830	-1,8%	2,70	,026

Tabelle 5-9: Maßzahlen der Verteilung der Differenzen von R^2, CPM und BIC im Rahmen des Vergleichs der Varianten 3 und 2 in der Hierarchie „Metabolische Erkrankungen“

a) Betroffenen-Konzept I

Verteilung der Differenz	Median	Mittelwert	Standard-abweichung	Variations-koeffizient	Schiefe	Standard-fehler der Schiefe
$R^2_3 - R^2_2$	,0331	,0331	,000363	1,1%	-1,602	,026
$CPM_3 - CPM_2$	,0087	,0087	,000029	0,3%	-,094	,026
$BIC_3 - BIC_2$	-,0619	-,0617	,001066	-1,7%	1,30	,026

b) Betroffenen-Konzept III

Verteilung der Differenz	Median	Mittelwert	Standard-abweichung	Variations-koeffizient	Schiefe	Standard-fehler der Schiefe
$R^2_3 - R^2_2$	,542	,536	,0210	3,9%	-4,3	,026
$CPM_3 - CPM_2$	,090	,090	,0021	2,4%	-,427	,026
$BIC_3 - BIC_2$	-,8345	-,8227	,040291	-4,9%	3,44	,026

5.5 Erforderliche Anzahl n von Wiederholungen der Bewertungsstichprobe

Da man mit dem Mikroskop-Design Neuland betreten hat, wurde für die Erprobung zunächst ein sehr hohes n für das Resampling vorgesehen, um die Frage, wie hoch n zu wählen ist, an einer ausreichend dimensionierten Datenbasis prüfen zu können. Zu diesem Zweck untersuchen wir im Folgenden einerseits die Größe des Mikroskopeffekts[8] und anderseits die Möglichkeiten der statistischen Absicherung der Mittelwerte der Differenzen. Vorab ist zu berichten, das die Vorzeichen der Differenzen sich für alle Vergleiche in jeder der insgesamt 9.000 Bewertungsstichproben als konstant herausgestellt haben (entweder immer „+“ oder immer „–“), was bereits als ein Hinweis auf die außerordentliche Stabilität des Designs gewertet werden muss.

5.5.1 Mikroskopeffekt

Wenn man das weite Betroffenen-Konzept zugrunde legt, ist der Mikroskopeffekt relativ klein und könnte fast besser als Lupeneffekt bezeichnet werden. Die Vergrößerung variiert zwischen 3,2 (Minimum) und 43,4 (Maximum). Anders, wenn das enge Betroffenen-Konzept verfolgt wird. Nun finden wir Werte zwischen 23 (Minimum) und 6.625 (Maximum), wobei letzteres mit einem Vorzeichenwechsel einhergeht (vgl. die Tabellen 5.10 bis 5.12).

Vergleicht man die Hierarchien untereinander, so ändert sich die Rangliste je nach dem zugrunde gelegten Betroffenen-Konzept und der betrachten Maßzahl. Betrachten wir zunächst die Verhältnisse für die weite Operationalisierung von „betroffen“. Hinsichtlich der R^2-Differenzen führt dann die Hierarchie „Metabolische Erkrankungen“ und die Hierarchie „Erkrankungen der Lunge“ bildet das Schlusslicht. Letztere steht aber hinsichtlich der CPM-Differenzen vorn. Wenn die enge Operationalisierung zugrunde gelegt wird, so führt die Hierarchie „Neubildungen“ die Liste für beide Differenzen an und die Hierarchie „Metabolische Erkrankungen“ bildet das Schlusslicht.

Die eigentliche Überraschung bietet aber die zu beobachtende Stabilität des Mikroskopeffektes gegenüber einer Variation des Stichprobenumfangs. Angesichts der guten Übereinstimmung der Effekte für n=1 mit entsprechenden Werten für n=9.000 (mit Ausnahme des Vergleichs der Varianten 3 und 2 in der Hierarchie „Neubildungen“) kann man sich fast zu der Aussage hinreißen lassen, es hätte der Wiederholungsziehungen der Bewertungsstichprobe gar nicht bedurft. In der Tat würde man für jeden der sechs Vergleiche, geschätzt auf der Basis der Differenzen der ersten Bewertungsstichprobe, zum gleichen Ergebnis kommen, wie auf der Basis der mittleren Differenzen aus 9.000 Bewertungsstichproben. Dies spricht für eine ganz außerordentliche Stabilität des Mikroskop-Designs.

[8] Es sei an die Definition des Mikroskopeffekts erinnert: Verhältnis der mittleren Differenz, berechnet aus den Bewertungsstichproben, zu der analogen Differenz, berechnet aus dem vollen Datenbestand.

Die Übereinstimmung der Mikroskopeffekte, berechnet aus n=100 Bewertungsstichproben, mit denjenigen, berechnet aus n=9.000 Bewertungsstichproben ist immens gut. Bei den kleineren Effekten finden wir exakte Übereinstimmung, bei den großen sind die relativen Abweichungen für die CPM-Differenzen maximal ca. 1%, für die R^2-Differenzen maximal ca. 5%.

Tabelle 5-10: Mikroskopeffekt bei Vergleichen in der Hierarchie „Erkrankungen der Lunge“ für verschiedene Stichprobenumfänge n

a) Vergleich der Varianten 2 und 1

	Betroffenen-Konzept I		Betroffenen-Konzept II	
n	$R^2_2 - R^2_1$	$CPM_2 - CPM_1$	$R^2_2 - R^2_1$	$CPM_2 - CPM_1$
1	3,2	43,2	341	-3.523
100	3,2	43,3	317	-3.354
9.000	3,2	43,4	317	-3.357

b) Vergleich der Varianten 3 und 1

	Betroffenen-Konzept I		Betroffenen-Konzept III	
n	$R^2_3 - R^2_1$	$CPM_3 - CPM_1$	$R^2_3 - R^2_1$	$CPM_3 - CPM_1$
1	4,3	6,2	315	597
100	4,4	6,2	302	581
9.000	4,4	6,2	303	582

Tabelle 5-11: Mikroskopeffekt bei Vergleichen in der Hierarchie „Neubildungen“ für verschiedene Stichprobenumfänge n

a) Vergleich der Varianten 2 und 1

	Betroffenen-Konzept I		Betroffenen-Konzept II	
n	$R^2_2 - R^2_1$	$CPM_2 - CPM_1$	$R^2_2 - R^2_1$	$CPM_2 - CPM_1$
1	4,9	6,6	514	-6.474
100	4,7	6,5	549	-6.625
9.000	4,7	6,5	541	-6.597

b) Vergleich der Varianten 3 und 2

	Betroffenen-Konzept I		Betroffenen-Konzept III	
n	$R^2_3 - R^2_2$	$CPM_3 - CPM_2$	$R^2_3 - R^2_2$	$CPM_3 - CPM_2$
1	4,7	5,6	928	2.363
100	4,8	5,6	2.340	2.579
9.000	4,8	5,6	2.226	2.550

Tabelle 5-12: Mikroskopeffekt bei Vergleichen in der Hierarchie „Metabolische Erkrankungen“ für verschiedene Stichprobenumfänge n

a) Vergleich der Varianten 2 und 1

	Betroffenen-Konzept I			Betroffenen-Konzept II		
n	$R^2_2 - R^2_1$	$CPM_2 - CPM_1$	$BIC_2 - BIC_1$	$R^2_2 - R^2_1$	$CPM_2 - CPM_1$	$BIC_2 - BIC_1$
1	9,0	21,4	10,9	22,8	48,1	26,8
100	8,9	21,3	10,7	22,6	47,5	26,5
9.000	8,9	21,3	10,8	22,6	47,5	26,5

b) Vergleich der Varianten 3 und 2

	Betroffenen-Konzept I			Betroffenen-Konzept II		
n	$R^2_3 - R^2_2$	$CPM_3 - CPM_2$	$BIC_3 - BIC_2$	$R^2_3 - R^2_2$	$CPM_3 - CPM_2$	$BIC_3 - BIC_2$
1	9,0	20,1	12,5	122	189	130
100	8,9	20,0	12,4	144	207	166
9.000	8,9	20,0	12,4	144	207	165

5.5.2 Variationskoeffizienten und statistische Absicherung der mittleren Differenzen.

Mit einer Ausnahme sind die Variationskoeffizienten der Differenzen klein bis sehr klein (vgl. die Tabellen 5.13 bis 5.15). Die Ausnahme bildet der Vergleich der Variante 3 mit der Variante 2 in der Hierarchie „Neubildungen“ unter Zugrundelegung des engen Betroffenen-Konzepts. Dies liegt daran, dass die einzelne Bewertungsstichprobe für diesen Vergleich einen besonders kleinen Stichprobenumfang besitzt (verursacht durch ein extrem kleines N_B=47, vgl. Tabelle 4-2).

Da das enge Betroffenen-Konzept generell zu kleineren Bewertungsstichproben führt, als das weitere, sind die zugehörigen Variationskoeffizienten dementsprechend größer.

Entscheidend ist aber, dass sich die Variationskoeffizienten nur wenig voneinander unterscheiden, wenn man die Verteilung der Differenzen einerseits aus nur n=100 n oder anderseits aus n=9.000 Bewertungsstichproben berechnet.

Insbesondere können bei Verfolgung des Mikroskop-Designs auch wenn man etwa nur n=100 Ziehungen der Bewertungsstichproben vorsieht, alle mittleren Differenzen als statistisch signifikant von Null verschieden erkannt werden, denn selbst der maximale Variationskoeffizient von 34% weist darauf hin, dass die Standardabweichung der zugehörigen Verteilung nur ca. ein Drittel des Mittelwertes beträgt. Zur Berechnung eines Konfidenzintervalles wird diese noch durch die Wurzel aus dem Stichprobenumfang (für n=100 also durch 10) dividiert.

Tabelle 5-13: Variationskoeffizient der Differenzen bei Vergleichen in der Hierarchie „Erkrankungen der Lunge“ für zwei Stichprobenumfänge

a) Vergleich der Varianten 2 und 1

	Betroffenen-Konzept I		Betroffenen-Konzept II	
n	$R^2_2 - R^2_1$	$CPM_2 - CPM_1$	$R^2_2 - R^2_1$	$CPM_2 - CPM_1$
100	1,8%	0,2%	8,0%	1,9%
9.000	1,8%	0,2%	7,1%	1,9%

b) Vergleich der Varianten 3 und 1

	Betroffenen-Konzept I		Betroffenen-Konzept III	
n	$R^2_3 - R^2_1$	$CPM_3 - CPM_1$	$R^2_3 - R^2_1$	$CPM_3 - CPM_1$
100	2,2%	0,8%	7,1%	1,9%
9.000	1,9%	0,8%	6,8%	2,0%

Tabelle 5-14: Variationskoeffizient der Differenzen bei Vergleichen in der Hierarchie „Neubildungen“ für zwei Stichprobenumfänge

a) Vergleich der Varianten 2 und 1

	Betroffenen-Konzept I		Betroffenen-Konzept II	
n	$R^2_2 - R^2_1$	$CPM_2 - CPM_1$	$R^2_2 - R^2_1$	$CPM_2 - CPM_1$
100	5,5%	0,5%	5,5%	2,0%
9.000	4,6%	0,5%	7,8%	2,1%

b) Vergleich der Varianten 3 und 2

	Betroffenen-Konzept I		Betroffenen-Konzept III	
n	$R^2_3 - R^2_2$	$CPM_3 - CPM_2$	$R^2_3 - R^2_2$	$CPM_3 - CPM_2$
100	1,9%	0,3%	32,4%	7,6%
9.000	1,5%	0,2%	34,0%	7,9%

Tabelle 5-15: Variationskoeffizient der Differenzen bei Vergleichen in der Hierarchie „Metabolische Erkrankungen“ für zwei Stichprobenumfänge

a) Vergleich der Varianten 2 und 1

	Betroffenen-Konzept I			Betroffenen-Konzept II		
n	$R^2_2 - R^2_1$	$CPM_2 - CPM_1$	$BIC_2 - BIC_1$	$R^2_2 - R^2_1$	$CPM_2 - CPM_1$	$BIC_2 - BIC_1$
100	1,0%	0,3%	1,4%	0,9%	0,7%	1,3%
9.000	1,1%	0,3%	1,5%	1,3%	0,7%	1,8%

b) Vergleich der Varianten 3 und 2

	Betroffenen-Konzept I			Betroffenen-Konzept III		
n	$R^2_3 - R^2_2$	$CPM_3 - CPM_2$	$BIC_3 - BIC_2$	$R^2_3 - R^2_2$	$CPM_3 - CPM_2$	$BIC_3 - BIC_2$
100	1,0%	0,3%	1,6%	2,7%	2,5%	3,7%
9.000	1,1%	0,3%	1,7%	3,9%	2,4%	4,9%

5.5.3 Zwischenresümee

Das Mikroskop-Design zeigt ein außergewöhnliches und so nicht erwartetes stabiles Verhalten. Schon mit nur einer Bewertungsstichprobe (ohne Wiederholungsziehung) würde man in allen sechs exemplarisch betrachteten Vergleichen zum gleichen Ergebnis kommen, wie auf der Basis der mittleren Differenzen, berechnet aus 9.000 Wiederholungsziehungen.

Um das Verfahren auch für alle zukünftigen Vergleiche anwendbar zu gestalten, sollten aber Wiederholungsziehungen vorgesehen werden, damit die empirische Verteilung der Differenzen in den Bewertungsprozess mit einbezogen werden kann. Die Zahl der Wiederholungsziehungen braucht nach den Erfahrungen aus den exemplarisch untersuchten sechs Vergleichen nicht sehr hoch zu sein. Da der Aufwand zur Ziehung der Wiederholungsstichproben vom Bundesversicherungsamt für ca. n=100 Wiederholungen als „vertretbar“ eingeschätzt wird, dürfte n=100 eine zweckmäßige Festlegung der Wiederholungszahl darstellen. Natürlich ist die Frage erlaubt, ob nicht vielleicht auch n=90, oder besser n=110 gewählt werden sollte. Zur Beantwortung dieser Frage müssen wir uns mit der konkreten Festlegung auf das Phänomen der „digital preference“ berufen.

5.6 Erörterung der sechs Bewertungsentscheidungen

In diesem Abschnitt stützen wir die Entscheidungen, welche Ausgestaltungsvariante in den drei betrachten Hierarchien jeweils zu bevorzugen ist, auf die Differenzen von R^2, MAPE und r, der Korrelation zwischen den tatsächlichen und den vorhergesagten Ausgaben (sowie ggf. auf die Differenz von BIC). Die Gründe hierfür sind in Abschnitt 3.2. ausführlich erörtert worden. Insbesondere wird MAPE anstelle von CPM verwendet, weil beide Maßzahlen in der hier diskutierten Anwendung stets zur gleichen Entscheidung führen, der mittlere Vorhersagefehler (in Euro) aber anders als die dimensionslose Verhältniszahl CPM eine unmittelbar greifbare und anschauliche Bewertung vermittelt.

5.6.1 Überblick

In den folgenden Tabellen werden die Mittelwerte der Differenzen der o. g. drei Maßzahlen aus n=100 Bewertungsstichproben des Mikroskop-Designs (MD) den Werten der Differenzen, berechnet aus dem vollen Datenbestand gegenübergestellt. Analog wird mit den Bewertungsentscheidungen verfahren.

Aus den Tabellen 5.16 bis 5.18 geht hervor:

1. Bewertungen auf der Basis der Differenz der R^2-Werte führen stets zum gleichen Ergebnis, ungeachtet der Tatsache, ob man sie aus dem vollen Datenbestand oder als Mittelwerte aus 100 Bewertungsstichproben des Mikroskop-Designs berechnet.
2. Alle Bewertungen stimmen überein, ungeachtet der speziell verwendeten Maßzahl und der herangezogenen Datenquelle (voller Datenbestand oder Mittelwert aus 100 Bewertungsstichproben des Mikroskop-Designs), sofern nur das weite Betroffen-Konzept zugrunde gelegt worden ist.

3. Bei Zugrundelegung des engen Betroffenen-Konzepts gibt es zwei Ausnahmen von dieser generellen Übereinstimmung: Bewertungen die auf dem mittleren absoluten Vorhersagefehler oder der Korrelation zwischen den tatsächlichen und den vorhergesagten Ausgaben beruhen, fallen für die Vergleiche der Varianten 2 und 1 in den beiden Hierarchien „Erkrankungen der Lunge“ und „Neubildungen“ andersherum aus, als diejenigen, die auf R^2 basieren. Sie weichen auch ab von den „eigenen“ Bewertungen bei Zugrundelegung des weiten Betroffenen-Konzepts.
4. Für die Vergleiche in der Hierarchie „Metabolische Erkrankungen“ stimmen alle Bewertungen überein, welche Maßzahldifferenz man auch immer heranziehen mag und ungeachtet der zur Berechnung herangezogenen Datenquelle.

Es bleibt zu klären, welche Bewertung man im Fall abweichender Ergebnisse bevorzugt. Diese Frage ist ohne Bezug auf den konkreten Vergleich nicht einfach zu beantworten. Wir erörtern sie vertieft in den Abschnitten 5.6.2 und 5.6.3.

Tabelle 5-16: Maßzahldifferenzen und Bewertungsentscheidungen für Vergleiche in der Hierarchie „Erkrankungen der Lunge“

a) Vergleich der Varianten 2 und 1

	Betroffenen-Konzept I			Betroffenen-Konzept II		
Differenz	$R^2_2 - R^2_1$	$MAPE_2 - MAPE_1$	$r_2 - r_1$	$R^2_2 - R^2_1$	$MAPE_2 - MAPE_1$	$r_2 - r_1$
Voller Datensatz	-0,0000032	0,006 €	-0,0000033	-0,0000032	0,006 €	-0,0000033
MD, n=100	-0,0000104	0,44 €	-0,0000148	-0,0010210	-44,08 €	0,0025456
Bewertung						
Voller Datensatz	V 1 besser	V 1 besser	V 1 besser	V 1 besser	V 1 besser	V 1 besser
MD, n=100	V 1 besser	V 1 besser	V 1 besser	V 1 besser	V 2 besser	V 2 besser

b) Vergleich der Varianten 3 und 1

	Betroffenen-Konzept I			Betroffenen-Konzept III		
Differenz	$R^2_3 - R^2_1$	$MAPE_3 - MAPE_1$	$r_3 - r_1$	$R^2_3 - R^2_1$	$MAPE_3 - MAPE_1$	$r_3 - r_1$
Voller Datensatz	0,00023	-0,29 €	0,00024	0,00023	-0,29 €	0,00024
MD, n=100	0,00102	-3,05 €	0,00106	0,07057	-538,85 €	0,05581
Bewertung						
Voller Datensatz	V 3 besser	V 3 besser	V 3 besser	V 3 besser	V 3 besser	V 3 besser
MD, n=100	V 3 besser	V 3 besser	V 3 besser	V 3 besser	V 3 besser	V 3 besser

Tabelle 5-17: Maßzahldifferenzen und Bewertungsentscheidungen für Vergleiche in der Hierarchie „Neubildungen“

a) Vergleich der Varianten 2 und 1

	Betroffenen-Konzept I			Betroffenen-Konzept II		
Differenz	$R^2_2 - R^2_1$	$MAPE_2 - MAPE_1$	$r_2 - r_1$	$R^2_2 - R^2_1$	$MAPE_2 - MAPE_1$	$r_2 - r_1$
Voller Datensatz	0,0000050	-0,02 €	0,0000046	0,0000050	-0,024 €	0,0000046
MD, n=100	0,0000231	-0,35 €	0,000019	0,0027287	363,71 €	-0,0008104
Bewertung						
Voller Datensatz	V 2 besser	V 2 besser	V 2 besser	V 2 besser	V 2 besser	V 2 besser
MD, n=100	V 2 besser	V 2 besser	V 2 besser	V 2 besser	V 1 besser	V 1 besser

b) Vergleich der Varianten 3 und 2

	Betroffenen-Konzept I			Betroffenen-Konzept III		
Differenz	$R^2_3 - R^2_2$	$MAPE_3 - MAPE_2$	$r_3 - r_2$	$R^2_3 - R^2_2$	$MAPE_3 - MAPE_2$	$r_3 - r_2$
Voller Datensatz	0,000037	-0,04 €	0,000038	0,000037	-0,04 €	0,000038
MD, n=100	0,00018	-0,45 €	0,00018	0,085748	-98,7 €	0,03567
Bewertung						
Voller Datensatz	V 3 besser	V 3 besser	V 3 besser	V 3 besser	V 3 besser	V 3 besser
MD, n=100	V 3 besser	V 3 besser	V 3 besser	V 3 besser	V 3 besser	V 3 besser

Tabelle 5-18: Maßzahldifferenzen und Bewertungsentscheidungen für Vergleiche in der Hierarchie „Metabolische Erkrankungen“

a) Vergleich der Varianten 2 und 1

	Betroffenen-Konzept I			Betroffenen-Konzept II		
Differenz	$R^2_2 - R^2_1$	$MAPE_2 - MAPE_1$	$r_2 - r_1$	$R^2_2 - R^2_1$	$MAPE_2 - MAPE_1$	$r_2 - r_1$
Voller Datensatz	0,018	-2,05 €	0,018	0,018	-2,05 €	0,018
MD, n=100	0,156	-117,32 €	0,128	0,397	-274,14 €	0,355
Bewertung						
Voller Datensatz	V 2 besser	V 2 besser	V 2 besser	V 2 besser	V 2 besser	V 2 besser
MD, n=100	V 2 besser	V 2 besser	V 2 besser	V 2 besser	V 2 besser	V 2 besser

Differenz	$BIC_2 - BIC_1$	$BIC_2 - BIC_1$
Voller Datensatz	-0,023	-0,023
MD, n=100	-0,249	-0,615
Bewertung		
Voller Datensatz	V 2 besser	V 2 besser
MD, n=100	V 2 besser	V 2 besser

b) Vergleich der Varianten 3 und 2

	Betroffenen-Konzept I			Betroffenen-Konzept III		
Differenz	$R^2_3 - R^2_2$	$MAPE_3 - MAPE_2$	$r_3 - r_2$	$R^2_3 - R^2_2$	$MAPE_3 - MAPE_2$	$r_3 - r_2$
Voller Datensatz	0,004	-0,95 €	0,004	0,004	-0,95 €	0,004
MD, n=100	0,033	-50,87 €	0,024	0,539	-554,92 €	0,524
Bewertung						
Voller Datensatz	V 3 besser	V 3 besser	V 3 besser	V 3 besser	V 3 besser	V 3 besser
MD, n=100	V 3 besser	V 3 besser	V 3 besser	V 3 besser	V 3 besser	V 3 besser

Differenz	$BIC_3 - BIC_2$	$BIC_3 - BIC_2$
Voller Datensatz	-0,0050	-0,0050
MD, n=100	-0,0616	-0,8292
Bewertung		
Voller Datensatz	V 3 besser	V 3 besser
MD, n=100	V 3 besser	V 3 besser

5.6.2 Vergleich der Varianten 2 und 1 in der Hierarchie „Erkrankungen der Lunge“

In der folgenden Tabelle sind die beiden zu vergleichenden Hierarchievarianten hinsichtlich der Anpassungsgüte dokumentiert. Neben dem Korrelationskoeffizienten r zwischen den tatsächlichen und den vorhergesagten Ausgaben wurde auch dessen Quadrat r^2 in die Tabelle aufgenommen, weil dieser nach der Theorie mit R^2 übereinstimmt, sofern die Berechnungen auf der Basis des gleichen Datensatzes erfolgen, an dem das Regressionsmodell kalibriert worden ist. Allerdings finden wir im Teil a) der Tabelle theoriewidrig kleine Abweichungen zwischen R^2 und r^2. Das liegt an den in Abschnitt 2.2 erörterten Abweichungen des im Morbi-RSA eingesetzten Regressionsverfahrens vom OLS-Modell und spielt für die weitere Argumentation keine Rolle. Darüber hinaus wird aus den Maßzahlen, die auf dem vollen Datenbestand berechnet worden sind, deutlich, dass die Unterschiede zwischen den beiden Varianten was die Anpassungsgüte betrifft, verschwindend klein sind. So unterscheiden sich die durchschnittlichen absoluten Vorhersagefehler pro Versichertem des Gesamtbestandes um nur sechs Hundertstel eines Cent! Von daher könnte man guten Gewissens beide Varianten im Hinblick auf die Modellanpassung für gleich gut halten.

Tabelle 5-19: Maßzahlen und ihre Differenzen für die Varianten 2 und 1 der Hierarchie „Erkrankungen der Lunge“

a) Berechnung aus dem vollen Datenbestand

	R^2	r^2	MAPE	r
Variante 2	0,2325475	0,2328709	1.647,295 €	0,4825669
Variante 1	0,2325507	0,2328741	1.647,289 €	0,4825703
Differenz	-0,0000032	-0,0000032	0,006 €	-0,0000033

b) Berechnung aus 100 Bewertungsstichproben, Betroffenen-Konzept I

	R^2	r^2	MAPE	r
Variante 2	0,233722	0,234775	2.666,47 €	0,484522
Variante 1	0,233733	0,234789	2.666,03 €	0,484537
Differenz	-0,000010	-0,000014	0,44 €	-0,000015

c) Berechnung aus 100 Bewertungsstichproben, Betroffenen-Konzept II

	R^2	r^2	MAPE	r
Variante 2	0,2155	0,2279	3.447,63 €	0,4768
Variante 1	0,2165	0,2254	3.491,71 €	0,4743
Differenz	-0,0010	0,0024	-44,08 €	0,0025

Betrachten wir nun die Daten des Mikroskop-Designs unter Zugrundelegung des weitgefassten Betroffenen-Konzepts, so wird der Unterschied des durchschnittlichen absoluten Vorhersagefehlers auf 44 Cent vergrößert (was auch nicht viel ist) und die Variante 1 liegt weiterhin vor der Variante 2. Erst bei Anwendung des engen Betroffenen-Konzepts kippt die Bewertung auf der Basis von MAPE um, und nun ist der mittlere absolute Vorhersagefehler in der Variante 1 um 44,08 € größer als in der Variante 2. Das steht im Widerspruch zu Rangfolge, die sich nach Maßgabe der R^2-Werte ergibt, stimmt aber

überein mit der Bewertung auf der Basis der Korrelation zwischen den tatsächlichen und den vorhergesagten Ausgaben.

Um den möglichen Ursachen für die unterschiedlichen Bewertungen auf den Grund zu gehen, ziehen wir die mit den einzelnen HMGs verbunden Zuschläge zu Rate (vgl. Tabelle 5-20).

Die vorhergesagten Ausgaben â ändern sich für einen Versicherten der DXG454 wenn er nicht, wie es Variante 1 vorschreibt, einer eigenständigen Gruppe zugeordnet, sondern gemäß Variante 2 in die Hierarchie der HMGs 107 bis 110 eingegliedert wird. Da die tatsächlichen Ausgaben a von den Unterschieden in der Ausgestaltung unberührt bleiben, ändert sich also auch das zugehörige Residuum e = a-â..

Tabelle 5-20: Regressionskoeffizienten Beta (Zuschläge) für die Varianten 1 und 2 der Hierarchie „Erkrankungen der Lunge“

		Variante 1		Variante 2	
HMG	**Bezeichnung**	**Beta**	**N**	**Beta**	**N**
HMG107	Mukoviszidose	15.959,21 €	462	16.083,39 €	462
HMG108	Status asthmaticus [Alter>17Jahre], postinflammatorische Lungenfibrose	2.228,10 €	4.042	2.265,31 €	4.042
HMG109	COPD/Emphysem [Alter>17Jahre]	847,11 €	194.294	850,51 €	193.152
HMG110	Chronische obstruktive Brocnchitis [Alter<18Jahre]	326,58 €	3.502	318,29 €	3.497
HMG111	Aspiration und näher bezeichnete bakterielle Pneumonien	5.197,20 €	2.828	5.205,82 €	2.828
HMG112	Sonstige Pneumonien, Pleuritis, Empyem, Lungenabszess, pulmonale Insuffizienz	977,76 €	22.190	979,65 €	22.190
DXG454	Bronchiektasen	1.408,39 €	1.753	1.832,93 €	1.604

Dabei kann dieses im Absolutbetrag sowohl größer als auch kleiner werden, wie folgende Überlegung zeigt: Ein Versicherter, der keiner der HMGs der Hierarchie zugeordnet ist, sondern nur der DXG454, erhält in der Variante 2 ein höheres â zugewiesen, da dieses sich aus der Summe aller Zuschläge des Modells ergibt, und ein Summand dieser Summe (das Beta der DXG 454) in der Variante 2 größer ausfällt, als in Variante 1 (vgl. Tabelle 5-20). Unterstellt man, dass die tatsächlichen Ausgaben für die DXG454 zugeordneten Versicherten höher sind als die vorhergesagten[9], so rückt â daher näher an a heran und der Absolutbetrag des Residuums wird kleiner. Damit erhält der mittlere absolute Vorhersagefehler MAPE einen Summanden, der in der Variante 2 kleiner ausfällt, als in der Variante 1. Das gleiche gilt aber auch wenn wir diesen Term quadrieren. Wir erhalten dann einen Summanden der Residualvarianz, die im Zähler des Bruchs steht, der bei der Berechnung von R^2 von 1 abgezogen wird.

[9] Dies ist plausibel, weil das Modell am vollen Datensatz kalibriert wurde und das Regressionsverfahren eine Ausgleichsrechnung ist, so dass für eher teuere Versicherte gilt, dass die vorhergesagten tendenziell unter den tatsächlichen Ausgaben liegen.

Wenn die DxG454 nur solchen Versicherten zugewiesen wäre, so würden wir daher sowohl auf der Basis von MAPE, als auch auf der Basis von R^2 die Variante 2 bevorzugen.

Nun finden wir aber auch Versicherte, die neben der DXG454 eine der HMGs 107, 108, 109 oder 110 zugewiesen wurden. Hier tritt der umgekehrte Effekt auf. Diese Versicherte erhalten in der Variante 1 höhere vorhergesagte Ausgaben als in der Variante 2, weil sich in der Variante 1 die Zuschläge addieren, während in der Variante 2, weil die DXG454 in die Hierarchie der genannten HMGs eingeordnet ist, nur einer der Zuschläge als Summand eingeht. Wenn die DXG454 nur solchen Versicherten zugewiesen wäre, würde man (mit analoger Argumentation wie im vorletzten Absatz schließend) sowohl auf der Basis von MAPE, als auch auf der Basis von R^2 die Variante 1 bevorzugen.

In Wirklichkeit gibt es aber beide Typen von Versicherten und daher stellen sich die zu beobachtenden Differenzen $R^2_2 - R^2_1$ und $MAPE_2 - MAPE_1$ als Ergebnis einer komplexen Bilanzierung der gegenläufigen Effekte auf die Summenterme ein.

Da durch das Quadrieren der einzelnen Summanden die Unterschiede vergrößert werden, birgt eine Bewertung auf der Basis der R^2-Differenzen im Vergleich zu einer Bewertung auf der Basis der MAPE-Differenzen vor diesem Hintergrund die größere Gefahr von Instabilität in sich. Dies entspricht der größeren Empfindlichkeit von R^2 im Vergleich zu CPM gegenüber statistischen Ausreißern.

Dass man bei Zugrundelegung des weiten Betroffenen-Konzepts auf der Basis von R^2 und vom MAPE dennoch zur gleichen Bewertung kommt, ist im Hinblick auf die folgende Überlegung plausibel

Die relevanten Summen bestehen aus zwei Termen, wobei der erste aus der Summe über die „Betroffenen“ und der zweite aus der Summe über die Mitglieder der Stichprobe aus den „Nichtbetroffenen“ besteht:

(20) $$\sum_{i=1}^{m} | \hat{a}_i - a_i | \;=\; \sum_{i=1}^{N_B} | \hat{a}_i - a_i | + \sum_{i=1}^{N_{NB}} | \hat{a}_i - a_i |$$

(21) $$\sum_{i=1}^{m} (\hat{a}_i - a_i)^2 \;=\; \sum_{i=1}^{N_B} (\hat{a}_i - a_i)^2 + \sum_{i=1}^{N_{NB}} (\hat{a}_i - a_i)^2$$

Wird nun das weite Betroffenen-Konzept verfolgt, so enthält der erste Term jeweils 219.181 Summanden von Versicherten, die einer HMG der Hierarchie zugeordnet, aber nicht in die Unterschiede zwischen den beiden Varianten involviert sind (vgl. Tabelle 4-1). Die mittleren Absolutbeträge der Residuen dieser (vergleichsweise teuren) Versicherten sind tendenziell größer als sie wären, wenn die Versicherten als Zufallsstichprobe aus den Versicherten gezogen würden, die keiner HMG der Hierarchie zugeordnet sind. Die auf diese Versicherten entfallene Teilsummen von (20) bzw. (21) liefern daher jeweils den Löwenanteil des ersten Terms und können daher die oben beschriebenen Unterschiede maskieren, so dass die Bewertungen auf der Basis der beiden Maßzahlen übereinstimmen.

Zusammenfassend kann man sagen, dass die Anpassung des Modells in beiden Varianten der Hierarchie „Erkrankungen der Lunge“ in etwa gleich ist, so dass es kaum einen Unterschied macht, ob man sich für die eine oder die andere Variante entscheidet. Wenn man aber doch eine Variante vor der anderen auszeichnen möchte, sollte man dies auf der Basis der robusteren Maßzahl MAPE tun und das enge Betroffenen-Konzept zugrunde legen.

5.6.3 Vergleich der Varianten 2 und 1 in der Hierarchie „Neubildungen“

Die Verhältnisse in der Hierarchie „Neubildungen“ sind, was die Modellanpassung in Bezug auf die Varianten 1 und 2 betrifft denen in der gerade diskutierten Hierarchie „Erkrankungen der Lunge“ sehr ähnlich. Auch hier sind die Unterschiede in der Anpassung winzig (der mittlere absolute Vorhersagefehler, berechnet aus dem vollen Datenbestand beträgt gerade mal 3 Cent). Wenn man eine Variante vor den anderen auszeichnen möchte, so spricht der MAPE-Ansatz unter Zugrundelegung des engen Betroffenen-Konzepts dafür, dass man der Variante 1 den Vorzug gibt. Diese Entscheidung wird auch auf der Basis der Korrelation zwischen den vorhergesagten und den tatsächlichen Ausgaben bestätigt.

Auf die Einbeziehung der Betas im Rahmen einer vertiefenden Interpretation verzichten wir bei dieser Hierarchie, weil die Unterschiede der Varianten durch Umdefinition der HMGs bzw. Verschiebungen zwischen den HMGs entstehen.

Tabelle 5-21: Maßzahlen und ihre Differenzen für die Varianten 2 und 1 der Hierarchie „Neubildungen“

a) Berechnung aus dem vollen Datenbestand

	R^2	r^2	MAPE	r
Variante 2	0,2325327	0,2328555	1.647,29 €	0,4825510
Variante 1	0,2325277	0,2328510	1.647,32 €	0,4825464
Differenz	0,0000050	0,0000045	-0,024 €	0,0000046

b) Berechnung aus 100 Bewertungsstichproben, Betroffenen-Konzept I

	R^2	r^2	MAPE	r
Variante 2	0,227148	0,228213	3.568,92	0,477701
Variante 1	0,227125	0,228195	3.569,27	0,477682
Differenz	0,000023	0,000019	-0,35 €	0,000019

c) Berechnung aus 100 Bewertungsstichproben, Betroffenen-Konzept II

	R^2	r^2	MAPE	r
Variante 2	0,1746	0,17574	3.817,35 €	0,41901
Variante 1	0,1719	0,17642	3.453,64 €	0,41982
Differenz	0,0027	-0,00068	363,71 €	-0,00081

5.7 Resümee

Das Mikroskop-Design kann als ein zweckmäßiger Ansatz beschrieben werden, um die Anpassungsunterschiede herauszuarbeiten, die sich für das Regressionsmodell des Morbi-RSA ergeben, wenn eine Krankheitshierarchie in zwei verschiedenen Ausgestaltungsvarianten in das Versichertenklassifikationsmodell aufgenommen werden kann.

In der Erprobung hat sich gezeigt, dass für das Resampling ein moderater Stichprobenumfang ausreicht, der im Hinblick auf die Auswertungsorganisation im Bundesversicherungsamt mit n=100 zweckmäßig gewählt sein dürfte.

Eine Überhöhung des Stichprobenumfangs der „Nichtbetroffenen“ gegenüber dem balancierten Design hat sich nicht als zweckmäßig erwiesen.

Auch kommt die Idee des Mikroskop-Designs am besten zum Tragen, wenn der Begriff des Betroffenen eng gefasst wird. Dennoch sollten die Berechnungen für zukünftige Bewertungen stets auch das weit gefasste Betroffenen-Konzept einbeziehen, weil Übereinstimmungen oder Abweichungen der Bewertung zusätzliche Interpretationsspielräume eröffnen. Schließlich sehen die Krankenkassen der Frage, welche Ausgestaltungsvariante einer Hierarchie für den Morbi-RSA implementiert werden sollte, naturgemäß nicht neutral, sondern interessengeleitet. Sie werden diejenige Variante bevorzugen, die ihnen eine höhere Zuweisung aus dem Gesundheitsfond verspricht. In dem notwendigen, in Form von Anhörungen gepflegten Dialog zwischen dem Bundesversicherungsamt auf der einen und den Krankenkassen und ihren Verbänden auf der anderen Seite können die zusätzlich eröffneten Interpretationsspielräume fruchtbar genutzt werden.

Der mittlere absolute Vorhersagefehler (MAPE) ist für die hier zu erörternden Zwecke der Messung von Anpassungsgüte hervorragend geeignet. Der Fehler wird in Euro pro Versichertem quantifiziert und ist damit unmittelbar verständlich. Außerdem gilt: Die Bewertungsentscheidungen auf der Basis von MAPE sind immer gleichlaufend mit denen auf der Basis von CPM, weil der Nenner der Verhältniszahl CPM für alle Ausgestaltungsvarianten der gleiche ist. MAPE misst die Anpassungsgüte robuster als R^2 und ist daher dem Bestimmtheitsmaß für die hier diskutierten Zwecke im Zweifel vorzuziehen.

Wenn man die Konsistenz der Entscheidungen bei der Ausgestaltung einer Hierarchie über mehrere Jahre prüfen will, so ist zu beachten, dass sich die Nenner in verschiedenen Jahren im Allgemeinen voneinander unterscheiden. Für Untersuchungen im Längsschnitt sollte die absolute Maßzahl MAPE zur Bewertung ihre Größe daher durch die relative Maßzahl CPM flankiert werden.

Obwohl MAPE und CPM die Anpassungsgüte robuster als R^2 messen, sollte auf R^2 als die tradierte Maßzahl nicht verzichtet werden, zumal in vielen Situationen die Bewertungen auf der Basis der R^2-Differenzen mit denen auf der Basis der MAPE-Differenzen übereinstimmen und man, wenn das nicht der Fall ist, aus der Abweichung möglicherweise zusätzliche Erkenntnisse ziehen kann.

Wenn sich zwei Ausgestaltungsvarianten in der Zahl der Prädiktoren unterscheiden (was bei den untersuchten Beispielen nur im Bereich der Hierarchie „Metabolische Erkrankungen“ der Fall war), so empfiehlt es sich, zusätzlich zu den anderen Maßzahlen auch das Bayessche Informationskriterium BIC zu berechnen, da es einen Strafterm für zusätzliche Prädiktoren enthält.

Stimmen die Varianten in der Zahl der Prädiktoren überein, so sind die Bewertungen auf der Basis von BIC gleichlaufend mit denjenigen auf der Basis von R^2, da die Maßzahl BIC auf dem Zähler des bei der Berechnung von R^2 abgezogenen Bruches basiert.

6 Literatur

Akaike H (1972): Information theory and an extension of the maximum likelihood principle. Proc. 2nd Int. Symp. on Information Theory, Supp. to Problems of Control and Information Theory, S. 267-281

Akaike H (1974): A new look at the statistical model identification. IEEE Trans. Auto. Control 19: 716-723

Bundesversicherungsamt (2010): Erläuterungen zum Entwurf der Festlegung von Morbiditätsgruppen, Zuordnungsalgorithmus, Regressionsverfahren und Berechnungsverfahren für den Jahresausgleich 2011 („Anhörungsdokument“), Bonn

Bundesversicherungsamt (2010): Erläuterungen zur Festlegung von Morbiditätsgruppen, Zuordnungsalgorithmus, Regressionsverfahren und Berechnungsverfahren für den Jahresausgleich 2011, Bonn

Cumming RB, Cameron BA (2002): A Comparative Analysis of Claims-based Methodes of Health Risk Assessment for Commercial Populations. A research study sponsored by the Society of Actuaries, May 24, Minneapolis

Greene WH (2008): Econometric Analysis. Pearson Prentice Hall, Upper Saddle River, New Jersey 07458

Harrel Jr. FE (2001): Regression Modeling Strategies with Applications to Linear Models. Logistic Regression, and Survival Analysis. Springer, New York

Hastie T, Tibshirani R, Friedman J (2001): The Elements of Statistical Learning – Data Mining, Inference and Prediction. Springer, New York

Kuha J (2004): AIC and BIC – Comparisons of Assumptions and Performance. Sociological Methods&Research, 33 (2): S. 188-229

Mallows CL (1973): Some Comments on CP. Technometrics 15 (4), S. 661–675.

Schwarz G (1978): Estimating the Dimension of a Model. In: Annals of Statistics. 2 (6) 1978, S. 461-464

Theil H (1971): Principles of Econometrics. Wiley, New York