Institut für Sozialmedizin, Epidemiologie
und Gesundheitssystemforschung

ISEG

# Expertise zur Eignung der Datengrundlage des Bewertungsausschusses zur Ermittlung der Veränderung der Morbiditätsstruktur gemäß § 87a Abs. 5 SGB V unter stichprobentheoretischen Gesichtspunkten

- im Auftrag des GKV-Spitzenverbandes -

Prof. Dr. Thomas Schäfer,

Oberuhldingen im August 2009

# Inhaltsverzeichnis

# Tabellenverzeichnis

# Zusammenfassung

Gegenstand der vorliegenden Expertise ist eine Untersuchung der Eignung der Datengrundlage des Bewertungsausschusses (die sog. 4KVen-Stichprobe) zur Ermittlung der Veränderung der Morbiditätsstruktur gemäß § 87a Abs. 5 SGB V.

Wenn die Datengrundlage des Bewertungsausschusses als Stichprobe angesehen wird, so handelt es sich hierbei um eine zweistufige Klumpenstichprobe. Auf der ersten Stufe erfolgt eine Auswahl von vier aus 17 Klumpen. Auf der zweiten Stufe wird eine einfache Zufallsstichprobe gezogen. Die Designeffekte einer Klumpenstichprobe mit so großen Klumpen, wie es die Versichertenkollektive der KVen darstellen, führt zu erheblichen merkmalsspezifischen Designeffekten. So ergibt eine Abschätzung, dass der Designeffekt des versichertenbezogenen Leistungsbedarfs auf der ersten Stufe über 200 beträgt. Um diesen Faktor ist die Varianz der Schätzung des durchschnittlichen Leistungsbedarfs aus dem 4KVen-Datenbestand größer als die Varianz der Schätzung aus einer einfachen Zufallsstichprobe mit gleich großer Versichertenzahl. Das Verhältnis der Standardabweichungen beträgt rund 15, um diesen Faktor ist das Konfidenzintervall breiter.

Es gibt Grund zu der Annahme, dass sich die Abschätzungen für den durchschnittlichen Leistungsbedarf auf die Schätzung der Veränderungsrate der Morbidität im Großen und Ganzen übertragen lassen. Daher ist zu befürchten, dass der Schätzfehler sehr viel größer ausfällt, als dies bei Schätzung aus einer einfachen Zufallsstichprobe der Fall wäre. Eine Simulation mit 10 Millionen von generierten Datensätzen hat ergeben, dass die allein auf die Klumpenstruktur des 4KVen-Datenbestandes zurückzuführende Schwankung der Veränderungsrate in der Worst-Case-Variante etwa ± 1,3 Prozentpunkte beträgt.

Es wäre daher besser, der Ermittlung der Veränderung der Morbiditätsstruktur gemäß § 87a Abs. 5 SGB V eine einfache Zufallsstichprobe zugrunde zu legen.

Sollte die Morbiditätsveränderungsrate dennoch aus dem 4KVen-Datenstand ermittelt werden, wäre das Institut des Bewertungsausschusses gut beraten, wenn es

1. die Kalibrierung des Regressionsmodells unter Berücksichtigung der Klumpenstruktur durchführt (z. B. mit Hilfe der Prozedur Surveyreg von SAS), weil die Schätzung der Beta-Gewichte dann mit korrekten Konfidenzintervallen versehen werden kann.
2. die tatsächliche Standardabweichung der Veränderungsrate durch Anwendung von Resamplingverfahren schätzt, um die Punktschätzung der Veränderungsrate mit einer Fehlerbreite absichern zu können.

Die mangelnde Schichtung der Grundgesamheit bei der „Auswahl“ der vier KVen geht über die aus der Klumpung resultierenden Mängel hinaus mit einer weiteren Vergrößerung der Varianz und der Gefahr des Verlustes an Repräsentativität einher.

# 1 Hintergrund und Vorgehensweise

Der Bewertungsausschuss hat gemäß § 87a Abs. 5 Satz 1 Nr. 2 SGB V ein Verfahren zur Bestimmung von Veränderungen der Morbiditätsstruktur zu beschließen, welches bei der Anpassung des Behandlungsbedarfs gemäß § 87a Abs. 4 SGB V angewendet wird.

Die Morbidität soll hierbei auf der Grundlage eines angepassten Klassifikationsverfahrens der DCG-Familie erfasst werden. Die Berechnungen auf der Basis eines definierten Kollektivs von Versicherten der gesetzlichen Krankenkassen werden vom Institut des Bewertungsausschusses durchgeführt.

Nach dem derzeitigen Stand ist der heranzuziehende Versichertenbestand definiert durch die Zugehörigkeit zur Wohnbevölkerung im PLZ-Bereich einer von vier KVen (Bremen, Niedersachsen, Nordrhein und Thüringen).

Der GKV-Spitzenverband hat grundsätzliche Zweifel, ob dieses Versichertenkollektiv für die Zwecke der Messung von Veränderung der Morbiditätsstruktur geeignet ist.

Mit der hier vorgelegten Expertise wird die Auswahl der vier 4 KVen unter stichprobentheoretischen Gesichtspunkten erörtert.

# 2 Verfahren zur Ermittlung der Veränderung der Morbiditätsstruktur

Auf der Grundlage des Datenbestandes der KVen soll die Veränderung der Morbiditätsstruktur in Deutschland ermittelt werden (und zwar voraussichtlich zwischen den Jahren 2007 und 2008). Für die Bewertung des Datenbestandes ist es wichtig zu wissen, in welcher Weise dies geschehen soll. Außerdem ist das Datengewinnungsverfahren von Interesse.

Der Datenbestand entsteht im ersten Schritt durch Zusammenführung („matching“) der pseudonymisierten Versichertendaten aller Versicherten der GKV mit Wohnsitz in einem der Organisationsgebiete der vier KVen mit den pseudonymisierten Abrechnungsdaten aller an der vertragsärztlichen Versorgung teilnehmenden Leistungserbringer mit Sitz in einem der Organisationsgebiete der vier KVen.

Danach werden im Rahmen einer Qualitätssicherung des Bewertungsausschusses die Daten verschiedener Versicherter und teilweise alle Versicherten ganzer Kassen markiert und bei späteren Verarbeitungsschritten nicht mehr berücksichtigt. Darüber hinaus werden Fehler in einzelnen Datensätzen, die bei Prüfung der Daten auf Plausibilität gefunden werden (z. B. negative Punktzahlen, mehr Versichertentage als das Quartal enthält, usw.), korrigiert

Schließlich wird eine Zufallsstichprobe aus dem so bereinigten Datenbestand gezogen, der dem Institut des Bewertungsausschusses für die weiter Bearbeitung übergeben wird.

Nach dem aktuellen Stand der Diskussion soll dann das o. g. Klassifikationsverfahren in ähnlicher Weise wie im Morbi-RSA in ein Regressionsmodell eingebettet und die Zuschläge (Beta-Gewichte) durch Kalibrierung in einem zeitgleichen oder prospektiven Ansatz gewonnen werden. Die Gewichte werden für die beiden Jahre, die der Veränderungsmessung zugrunde gelegt werden, festgeschrieben.

In einem zweiten Schritt wird die Summe der Zuschläge (oder der Relativgewichte) für jedes der beiden Jahre über alle Versicherten gebildet und durch die zugehörigen Versichertenjahre dividiert, um den Morbiditätsindex (MI) für das jeweilige Jahr zu erhalten.

Dividiert man dann den MI des Folgejahres durch den des Ausgangsjahres, zieht davon 1 ab und multipliziert das Ergebnis mit Hundert, so ergibt sich die prozentuale Veränderung des MI.

# 3 Klumpenstichprobe und Designeffekt

## 3.1 Abschätzung des Designeffekts für die Veränderungsrate

Der Datenbestand der vier KVen ist im Sinne der Stichprobentheorie keine Stichprobe, da er für den hier diskutierten Zweck nicht auf der Basis einer kriteriengesteuerten Stichprobenplanung gewonnen wurde. Man kann ihn aber im Sinne eines Best-Case-Szenarios so behandeln, als wäre er Resultat einer Stichprobenplanung, und kann dann die Rationalität, Qualität und Effizienz eines solchen Stichprobenplans kritisch hinterfragen.

In diesem Best-Case-Szenario würde es sich um eine zweistufige Klumpenstichprobe handeln, wobei die zweite Stufe als echte bzw. Pseudozufallsauswahl der Versicherten aus den in der ersten Stufe gezogenen Klumpen (= Versicherte mit Wohnsitz in einer der vier KV-Regionen) nicht das Problem darstellt.[1]

Es geht um die erste Stufe. Die Varianz der Schätzungen, die aus Daten einer Klumpenstichprobe gewonnen werden, ist erfahrungsgemäß größer (in Ausnahmefällen auch kleiner), als diejenige, die man für den gleichen Mittel- oder Anteilswert aus einer einfachen Zufallsstichprobe erhalten würde. Das Verhältnis dieser beiden Varianzen wird in der Stichprobentheorie als Designeffekt bezeichnet, der auf dem sog. Intraclasskorrelationskoeffizienten beruht. Letzterer kann als Maßzahl für die durchschnittliche Homogenität der Zusammensetzung der Klumpen hinsichtlich des betrachteten Merkmals angesehen werden. Er ist kein „normaler“ Korrelationskoeffizient. Vielmehr berechnet er sich aus einer Varianzanalysetabelle als eine geeignet normierte Differenz zwischen der Varianz zwischen den Klumpen und der Varianz innerhalb der Klumpen (beides in der Grundgesamtheit) relativ zur Gesamtvarianz.

Der Designeffekt ist selten kleiner als Eins, nämlich nur dann, wenn der Intraclasskorrelationskoeffizient negativ ist. Er ist um so größer, je homogener die Klumpen sind und je mehr Untersuchungseinheiten sie umfassen.

Obwohl der Klumpenumfang zwischen 570 Tsd. (KV Bremen) und 10,4 Mio. (KV Bayern) schwankt, betrachten wir die Verhältnisse zur Reduktion von Komplexität zunächst für eine einstufige Klumpenstichprobe mit fixer Klumpengröße M (wofür wir die mittleren Versichertenzahlen der 17 KVen einsetzen können, d.h. M = 4,1 Mio.).

Wir betrachten ein Merkmal Y der Versicherten (z.B. den Leistungsbedarf), dessen Mittelwert $\overline{Y}$ aus der Stichprobe geschätzt werden soll. Mit K = 17 und

$$\overline{Y}_{i\bullet} = \frac{1}{M}\sum_{j=1}^{M} Y_{ij} \qquad = \text{Mittelwert im i-ten Klumpen der Grundgesamtheit}$$

ergibt sich die folgende Varianzanalysetabelle (vgl. z.B. Schäfer, 2004):

1 Probleme, die sich aus den Bereinigungen der Daten ergeben könnten, bleiben im Folgenden unberücksichtigt.

| Variation | Quadratsumme (SQ) | Freiheits-grade (FG) | Mittlere Quadratsumme (MQ) |
|---|---|---|---|
| zwischen den Klumpen | $M\sum_{i=1}^{K}(\bar{Y}_{i\bullet}-\bar{Y})^2$ | K-1 | $MS_{zw}^{\;2} = \frac{M}{K-1}\sum_{i=1}^{K}(\bar{Y}_{i\bullet}-\bar{Y})^2$ |
| innerhalb der Klumpen | $\sum_{i=1}^{K}\sum_{j=1}^{M}(Y_{ij}-\bar{Y}_{i\bullet})^2$ | K(M-1) | $S_{in}^{\;2} = \frac{1}{K(M-1)}\sum_{i=1}^{K}\sum_{j=1}^{M}(Y_{ij}-\bar{Y}_{i\bullet})^2$ |
| Grundgesamtheit insgesamt | $\sum_{i=1}^{K}\sum_{j=1}^{M}(Y_{ij}-\bar{Y})^2$ | KM-1 | $S^{\;2}=\frac{1}{KM-1}\sum_{i=1}^{K}\sum_{j=1}^{M}(Y_{ij}-\bar{Y})^2$ |

Tabelle 3-1 Varianzanalysetabelle der einfachen Klumpenstichprobe mit konstanter Klumpengröße

Macht man die Korrektur der in der letzten Spalte von Tabelle 3-1 ausgewiesenen mittleren Quadratsummen auf Erwartungstreue wieder rückgängig, so erhält man:

$$\sigma_{zw}^{\;2} = \frac{K-1}{K} S_{zw}^{\;2}$$ (Varianz zwischen den Klumpen)

$$\sigma_{in}^{\;2} = \frac{M-1}{M} S_{in}^{\;2}$$ (Varianz innerhalb der Klumpen)

$$\sigma^2 = \frac{KM-1}{KM} S^2$$ (Gesamtvarianz)

Diese Varianzen sind verknüpft in der Streuungszerlegungsgleichung

(1) $$\sigma^2 = \sigma_{zw}^{\;2} + \sigma_{in}^{\;2}$$

und sie erlauben es, den Intraclasskorrelationskoeffizienten $\rho$ folgendermaßen zu berechnen:

(2) $$\rho = \frac{\sigma_{zw}^{\;2} - \sigma_{in}^{\;2}/(M-1)}{\sigma^2} .$$

Wenn man jetzt aus den K Klumpen, die zusammen die Grundgesamtheit konstituieren, k Klumpen zufällig auswählt (wobei alle Elemente dieser Klumpen in die Stichprobe kommen) und den Mittelwert $\bar{Y}$ durch das Stichprobenmittel $\bar{y}$ schätzt, dann ergibt sich als Verhältnis der Varianz von $\bar{y}$, berechnet aus der Klumpenstichprobe, zu der Varianz von $\bar{y}$, berechnet aus einer einfachen Zufallsstichprobe (ohne Zurücklegen) mit dem gleichen Stichprobenumfang kM, der sog. Designeffekt Deff in folgender Form (vgl. z. B. Sukhatme/Sukhatme,1970, S. 223 ff. oder Cochran. 1972, S. 285 f.):

(3) $$\mathrm{Deff} = (1+(M-1)\rho)$$

Die Formeln (2) und (3) werden vielfach auch für die einstufige Klumpenauswahl mit variabler Klumpengröße angewendet, wobei dann die mittlere Klumpengröße eingesetzt wird.

Geht man von einem positiven Intraclasskorrelationseffekt aus, was bei der Berechnung beispielsweise der mittleren Zuschläge für Versicherte in einer bestimmten Morbiditätskategorie wahrscheinlich ist, und was darüber hinaus von den Erfahrungen anderer Klumpenstichprobendesigns gedeckt wird, so ist daher eine Klumpenstichprobe, bei der die Grundgesamtheit in wenige Klumpen sehr großen Umfangs zerfällt, extrem ungünstig, da der Designeffekt linear mit der mittleren Klumpengröße wächst.

Der Intraclasskorrelationskoeffizient und der Designeffekt sind merkmalsspezifisch. Es gibt diese Größen nicht global, sondern man muss sie für jedes Merkmal neu berechnen.

In Folgenden sollen beide Größen für das Merkmal „Leistungsbedarf“ abgeschätzt werden. Der Grundgedanke für diese Abschätzung ist der folgende:

Der Mittelwert $\bar{y}$ hat das asymptotische Konfidenzintervall mit den Grenzen

(4) $$\bar{y} \pm 1{,}96\sqrt{\mathrm{Var}(\bar{y})}$$

wobei $\sqrt{\mathrm{Var}(\bar{y})}$ vom Stichprobendesign abhängt. In der einfachen Zufallsstichprobe (mit Stichprobenumfang kM ohne Zurücklegen) gilt

(5) $$\mathrm{Var}_{\mathrm{einf}}(\bar{y}) = (1\text{-}f)\frac{\sigma_G^2}{kM}, \qquad \text{mit} \quad f = \frac{kM}{KM} = \frac{k}{K} \quad \text{(Auswahlsatz)}$$

wenn k die Zahl der ausgewählten Klumpen ist und $\sigma_G^2$ die Varianz des untersuchten Merkmals in der Grundgesamtheit. Nach Angaben des GKV-Spitzenverbandes liegt der Variationskoeffizient (= Standardabweichung/Mittelwert) des Leistungsbedarfs der Versicherten zwischen 2,5 und 3, so dass die Standardabweichung um den entsprechenden Faktor größer ist als der Mittelwert. Für die nachstehende Rechnung verwenden den Faktor 2,75.

Mit dieser Information können wir die Varianz des Mittelwertes aus einer einfachen Zufallsstichprobe berechnen. Sie ist bei dem großen Stichprobenumfang sehr klein.

Da aber eine Klumpenstichprobe vorliegt, unterschätz die ausgewiesene Standardabweichung der einfachen Stichprobe $\sqrt{\mathrm{Var}_{\mathrm{einf}}(\bar{y})}$ die tatsächlich zu beobachtende Standardabweichung in der Klumpenstichprobe und das Konfidenzintervall ist entsprechend zu korrigieren:

(6) $$\bar{y} \pm 1{,}96\sqrt{\mathrm{Deff} \cdot \mathrm{Var}_{\mathrm{einf}}(\bar{y})}$$

Zur Ermittlung der tatsächlich zu beobachtenden Varianz von $\bar{y}$ dient die Überlegung, dass der Mittelwert der Grundgesamtheit vom Konfidenzintervall (6) in 95% der Anwendungen überdeckt wird. Man kann also den Mittelwert der Grundgesamtheit (der größer ist, als der Mittelwert in der Gruppe der vier KVen (vgl. Tabelle 3-2) mit einer gewissen Berechtigung als das Minimum der möglichen oberen Grenze des Konfidenzintervalls ansehen. Daraus erhält man eine Gleichung, die nach der Varianz von $\bar{y}$ bzw. nach dem Designeffekt aufgelöst werden kann:

(7) $$\overline{Y} = \overline{y} + 1{,}96\sqrt{\mathrm{Deff} \cdot \mathrm{Var}_{einf}(\overline{y})}$$

Löst man diese Gleichung nach $\sqrt{\mathrm{Deff}}$ auf, so erhält man:

(8) $$\sqrt{\mathrm{Deff}} = \frac{\overline{Y}\text{-}\overline{y}}{1{,}96 \cdot \sqrt{\mathrm{Var}_{einf}(\overline{y})}}$$

Die Berechungen für den Leistungsbedarf in Punkten auf der Basis der C4-Daten des Jahres 2007 sind in Tabelle 3-2 zusammengestellt.

Bezogen auf den Leistungsbedarf in Punkten liegt der Designeffekt bei 223 und das Verhältnis der Standardabweichungen bei rund 15.

Aus dem Designeffekt Deff und der mittleren Klumpengröße berechnet sich mit Hilfe der Formel (3) der Intraclasskorrelationskoeffizient in Bezug auf den Leistungsbedarf in Punkten zu $\rho = 0{,}000054$.

Die Aufblähung der Standardabweichung bei der Schätzung des Mittelwertes in der Grundgesamtheit führt zu Konfidenzintervallen, die um den gleichen Faktor verbreitert werden, d. h. die Schätzungen verlieren durch den Klumpeneffekt gewaltig an Präzision und Zuverlässigkeit. So berechnet sich das Konfidenzintervall für den Mittelwert des Leistungsbedarfs in Punkten der Grundgesamtheit – bei einem Durchschnittswert von 11.081 Punkten in der Gruppe der vier KVen – in der Form

11.081 ± 1,96 * 14,9 * 6,555 = 11.081 ± 191

gegenüber einem wesentlich kleineren Konfidenzintervall unter Annahme einer einfachen Zufallsstichprobe:

11.081 ± 1,96 * 6,555 = 11.081 ± 12,8

(vgl. Tabelle 3-2).

| | Leistungsbedarf in Punkten pro Versicherten |
|---|---|
| 4KVen | 11.081 |
| Alle KVen | 11.273 |
| Variationskoeffizient (Erfahrungswert) | 2,75 |
| Mittlere Klumpengröße (M) | 4.131.316 |
| 4xM | 16.525.264 |
| Standardabweichung des Mittelwertes (einfache Zufallsstichprobe) | 6,555 |
| Min. der oberen Grenze des Konfidenzintervalls (Klumpenstichprobe) | 11.273 |
| Min. der tatsächlichen Standardabweichung des Mittelwertes | 98,0 |
| Wurzel aus Designeffekt | 14,9 |
| Designeffekt | 223,3 |
| Intraclass-Korrelation | 0,000054 |

Tabelle 3-2 Abschätzung des Designeffekts und des Intraclasskorrelationskoeffizienten bei der Schätzung des Mittelwertes des Leistungsbedarfs in Punkten aus der „4KVen-Stichprobe“, Jahr=2007

Allerdings geht es bei der Ermittlung der Morbiditätsveränderungsrate nicht um einen Mittelwert, sondern um das Verhältnis der beiden Morbiditätsindices, berechnet für das Jahr 1 (im Zähler) und für das Jahr 0 im Nenner. Dies ist eine so komplexe Maßzahl, dass man keine fundierte quantifizierte Aussage über ihre Varianz bei Schätzung aus dem Datenbestand der 4KVen machen kann, ohne umfangreiche Berechnungen auf der Basis eben dieser Daten durchgeführt zu haben. Dennoch kann die Beziehung des gesuchten Designeffekts zu dem für das Versichertenmerkmal „Leistungsbedarf“ abgeschätzten in einigen Bemerkungen näher beschrieben werden:

(a) Man könnte den Morbiditätsindex MI für ein Jahr als Mittelwert des Versichertenmerkmals „Standardisierter Leistungsbedarf des Kalibrierungsjahrs“ interpretieren und die Formel (3) anwenden, wären nicht die summierten Beta-Gewichte $B_1$ bis $B_L$ (L sei die Anzahl der Morbiditätskategorien) selbst Zufallsvariable mit jeweils einer eigenen Varianz. In solchen Fällen hilft eine bedingte Betrachtungsweise. Seien $b_1$ bis $b_L$ konkrete Ausprägungen der Beta-Gewichte. Dann betrachtet man zunächst

$$\mathrm{Var}[MI|(B_1 = b_{1, \ldots,} B_L = b_L)],$$

d. h. die bedingte Varianz, gegeben $B_1 = b_{1, \ldots,}\ B_L = b_L$.

Diese wird, wenn der Morbiditätsindex aus dem 4KVen Datenbestand statt aus einer gleich großen einfachen Zufallsstichprobe geschätzt wird, zwar nicht ganz um den Faktor des abgeschätzten Designeffekts größer sein, da der Intraclasskorrelationseffekt beim standardisierten Leistungsbedarf vermutlich kleiner sein wird, als beim tatsächlich angefallenen, aber die Größenordnung des abgeschätzten Designeffekts bietet eine grobe Orientierung.

Die unbedingte Varianz von MI ergibt sich aus der bedingten dann in folgender Weise (s. z. B. Hinderer, 1972):

$$\text{(9)} \qquad \mathrm{Var}(MI) = E_B(\mathrm{Var}[MI|B]) + \mathrm{Var}_B(E[MI|B])$$

wobei mit „$E_B$“ und „$\mathrm{Var}_B$“ die Bildung des Erwartungswertes und der Varianz bzgl. der Verteilung von B gemeint ist

Der zweite Summand in (9) trägt der Variabilität der Schätzung der Regressionskoeffizienten Rechnung und lenkt den Blick auf die Designeffekte von Regressionskoeffizienten.

(b) Die Probleme, die im Rahmen der Berechnung von Varianzen und Designeffekten bei der Schätzung komplexer Modellparameter aus komplexen Stichproben auftreten, wurden schon 1974 in einer grundlegenden Arbeit von Kish und Frankel behandelt. Nach den Untersuchungen dieser beiden Stichprobentheoretikern fallen die Designeffekte von Regressionskoeffizienten nicht so hoch aus wie die von Mittelwerten, aber sie weisen eine Beziehung zu dem Designeffekt der Zielvariable des multiplen Regressionsmodells auf, so dass die Designeffekte von Regressionskoeffizienten für Zielvariablen mit hohen Designeffekten größer sind, als solche für Zielvariablen mit niedrigen Designeffekten.

(c) Für eine Klumpenstichprobe wurde der Designeffekt der Regressionskoeffizienten eines multiplen linearen Regressionsmodells von Holt/Scott (1981) konkret berechnet. Aus dieser Arbeit wird zunächst deutlich, dass für ein Regressionsmodell mit k Prädiktoren k+3 relevante Designeffekte betrachtet werden müssen, nämlich diejenigen der Zielvariablen, der k Prädiktoren, des Absolutglieds und der Residuen. Für das Modell der einfachen linearen Regression mit nur einem Prädiktor x ist der Designeffekt des Regressionskoeffizienten proportional dem Ausdruck

$$1+(M-1)\rho\rho_x,$$

wobei M die hier als konstant angenommene Klumpengröße bezeichnet, $\rho$ der Intraclasskorrelationskoeffizient der Residuen ist und $\rho_x$ derjenige des Prädiktors x. Da anzunehmen ist, dass $\rho\rho_x$ als Produkt zweier (in der Regel sehr kleiner) Korrelationskoeffizienten im Allgemeinen kleiner ist als $\rho_y$, werden die Beobachtungen von Kish und Frankel verständlich.

(d) Aus (9) berechnet sich der Designeffekt des Morbiditätsindex als gewichtete Summe der Designeffekte der beiden Summanden (die Gewichte sind die Anteile der beiden Terme, berechnet unter Zugrundlegung einer einfachen Zufallsstichprobe). Damit wird deutlich, dass die Varianz des Morbiditätsindexes MI für ein bestimmtes Jahr, berechnet aus dem 4KVen-Datenbestand zwar nicht um den abgeschätzten Faktor 223 größer ist, als wenn er aus einer einfachen Zufallsstichprobe berechnet wurde, aber doch vermutlich um einen nennenswerten Faktor im zwei- oder sogar dreistelligen Bereich, so dass größte Zweifel an der Genauigkeit der Schätzung des MI aus diesem Datenbestand angebracht sind.

(e) Die Veränderungsrate berechnet sich als Quotient der Morbiditätsindices zweier aufeinander folgender Jahre (Basisjahr: t = 0, Berichtsjahr: t = 1).[1] Die Varianzen von Quotienten sind nicht einfach zu berechnen. Jedoch können die Designeffekte nach Kish (1995) erfahrungsgemäß gut approximiert werden durch die Designeffekte der Differenzen, so dass man davon ausgehen kann, dass folgende Nährung gilt:

$$\text{(10)} \qquad \mathrm{Deff}\left(\frac{MI_1}{MI_0}\right) \approx \mathrm{Deff}(MI_1\text{-}MI_0) = \frac{\mathrm{Var}_{KL}(MI_1\text{-}MI_0)}{\mathrm{Var}_{einf}(MI_1\text{-}MI_0)}.$$

Diese Nährung gibt uns das Werkzeug in die Hand, nun auch den Designeffekt der Veränderungsrate abzuschätzen.

Unbeschadet der Art und Weise, wie die Stichprobe gezogen wurde, gilt:

$$\text{(11)} \qquad \mathrm{Var}(MI_1\text{-}MI_0) = \mathrm{Var}(MI_1) + \mathrm{Var}(MI_0) - 2\mathrm{Kov}(MI_1, MI_0)$$

1 Falls anstelle der Beta-Gewichte die Relativgewichte verwendet werden, so muss noch das Verhältnis der Durschnittwerte des Leistungsbedarfs der Versicherten im Bereich der vier KVen in den beiden Jahren als konstanter Faktor angebracht werden, der für die Berechnung des Designeffekts aber vernachlässigt werden kann.

Aus (10) und (11) ergibt sich

(12) $$\mathrm{Deff}\left(\frac{MI_1}{MI_0}\right) \approx \frac{\mathrm{Var}_{KL}(MI_1) + \mathrm{Var}_{KL}(MI_0) - 2\mathrm{Kov}_{KL}(MI_1, MI0_0)}{\mathrm{Var}_{einf}(MI_1) + \mathrm{Var}_{einf}(MI_0) - 2\mathrm{Kov}_{einf}(MI_1, MI_0)},$$

wobei die Varianzen und die Kovarianz im Zähler sich auf die tatsächliche Berechnung aus dem 4KVen-Datenbestand beziehen, und die im Nenner (durch den Index „einf“ angedeutet) auf Berechnung aus einer einfachen Zufallsstichprobe ohne Zurücklegen mit gleich vielen Versicherten. Es ist anzunehmen, dass die Kovarianz zwischen den beiden Morbiditätsindizes für Jahr 1 und Jahr 0 positiv und nicht ganz klein sein wird, da die Mittelwerte ja zu großen Teilen über die gleichen Versicherten berechnet werden.

Wenn man nun davon ausgeht, dass sich die drei Designeffekte bezogen auf die Varianzen von $MI_1$ und $MI_0$ und bezogen auf die Kovarianz zwischen beiden Morbiditätsindices nicht sehr stark unterscheiden (was plausibel ist) und ersetzt sie alle drei durch einen mittleren Designeffekt $\bar{D}$ so lässt sich der Bruch in (12) kürzen und wir erhalten, dass der Designeffekt für die Morbiditätsveränderungsrate annährend mit $\bar{D}$ übereinstimmt. Die Bedenken, die wir unter (d) gegenüber der Schätzung der Morbiditätsindices aus dem Datenbestand der vier KVen geäußert haben, übertragen sich auch auf die Schätzung der Veränderungsrate. Man muss mit einer Standardabweichung rechnen, die ein Mehrfaches der Standardabweichung bei Berechnung aus einer einfachen Zufallsstichprobe beträgt. Entsprechend breiter ist das Konfidenzintervall und entsprechen skeptisch muss man die Genauigkeit der Schätzung beurteilen.

### 3.2 Simulation der Auswirkungen der Klumpenstruktur auf die Veränderungsrate der Morbidität

Zur Illustration des potentiellen Effektes der Klumpenstruktur auf die Veränderungsrate der Morbidität sei nochmals auf die Daten aus Tabelle 3-2 zurückgegriffen. Darüber hinaus sei unterstellt, dass die Veränderungsrate aus dem Verhältnis des geschätzten Leistungsbedarfs (in Punkten) von Basisjahr (t=0) und Berichtsjahr (t=1) ermittelt wird und eine Veränderungsrate von 2,5% vorliegt. Um den Effekt quantifizieren zu können, sind in Tabelle 2-3 die Ergebnisse einer Monte-Carlo-Simulation zusammengestellt, mit deren Hilfe die Standardabweichung folgender Größe generiert wurde:

(13) $$\mathrm{VR} = \frac{(11081 \cdot 1{,}025 + \delta_1)}{(11081 + \delta_0)}$$

Hierbei sind $\delta_0$ und $\delta_1$ unabhängig voneinander erzeugte Zufallszahlen aus einer Normalverteilung mit Erwartungswert 0 und einer Standardabweichung für die einfache Zufallsstichprobe, wie sie sich aus Tabelle 3-2 ergibt (6,555).

Der Designeffekt der Klumpenstichprobe wurde in zwei Varianten simuliert:

a) In der Worst-Case-Variante wurde für den Designeffekt des Morbiditätsindexes MI unmittelbar derjenige des Leistungsbedarfs in Punkten aus Tabelle 3-2 übernommen.

b) In der realistischen Variante wurde zunächst unterstellt, dass der Designeffekt des standardisierten Leistungsbedarfs MI nur halb so groß ist, wie der in Tabelle 3-2 ausgewiesene. Sodann wurde mit Bezug auf Gleichung (9) angenommen, dass der Designeffekt des zweiten Summanden, mit dem die Variabilität der Schätzungen der Regressionskoeffizienten berücksichtigt wird, nur den zehnten Teil des Designeffekts des ersten Summanden beträgt. Schließlich wurde das Gewicht des Designeffekts des ersten Summanden mit 0,7, dasjenige des zweiten mit 0,3 angenommen.

Während bei Zugrundelegung einer einfachen Zufallsstichprobe sich für die Veränderungsrate eine Standardabweichung $s_{einf}$ = 0,091 ergibt, sind die aus der Monte-Carlo-Simulation abgeschätzten Standardabweichungen für die Veränderungsrate der Klumpenstichprobe mit $s_{Kl}$ = 0,822 (realistische Variante) bzw. $s_{Kl}$ = 1,363 (Worst-Case-Variante) sehr viel höher.

Es ergibt sich mithin aus dieser Simulation, dass die allein auf die Klumpenstruktur des 4KVen-Datenbestandes zurückzuführende Schwankung der Veränderungsrate in der Worst-Case-Variante ± 1,272 (=± (1,363-0,091)) Prozentpunkte beträgt. In der realistischen Variante sind es immerhin noch ± 0,731 Prozentpunkte.

| Kennzahlen der Veränderungsrate | Mittelwert (Prozent) | Standardabweichung (Prozentpunkte) | Variationskoeffizient |
|---|---|---|---|
| ohne Designeffekt (Deff = 1) für MI | 2,50 | 0,091 | 0,036 |
| mit Designeffekt Deff = 81,14 für MI (realistische Variante) | 2,45 | 0,822 | 0,336 |
| mit Designeffekt Deff = 223,3 für MI (Worst-Case-Variante) | 2,42 | 1,363 | 0,563 |

Tabelle 3-3 Monte-Carlo-Simulation mit 10 Millionen von generierten Datensätzen

### 3.3 Vernachlässigung der Schichtung

Eine weitere Schwäche des Datenbestandes ergibt sich aus der Vernachlässigung der Schichtung der Grundgesamtheit aller 17 KVen.

Man zieht eine Stichprobe aus zwei Gründen geschichtet: Einerseits soll die Varianz der Schätzungen verkleinert werden (infolge der Homogenität der Schichten). Anderseits soll, wenn nur wenige Einheiten ausgewählt werden, dadurch weitgehend sichergestellt werden, dass sich die Heterogenität der Grundgesamtheit auch in der Stichprobe wieder findet.

In den vier hier diskutierten KVen sind zwar Stadt- und Flächenstaaten vertreten, aber beispielsweise bleibt der Süden Deutschlands vollständig unberücksichtigt.

Im Rahmen einer soliden Stichprobenplanung hätte man, wenn (aus welchen Gründen auch immer) ein Klumpenstichprobendesign auf der Ebene der KVen als Nebenbedingung akzeptiert werden müsste, die 17 verschiedenen KVen auf Homogenität bzw. Heterogenität untersuchen müssen, um eine geeignete Schichtung nach Vergleichbarkeit im Leistungsbedarf und der Morbiditätsstruktur vornehmen zu können, aus der dann jeweils eine Anzahl von KVen per Zufall auszuwählen wären.

Es bleibt festzuhalten, dass der Mangel an Schichtung zwei Effekte hat:

1. Die Varianzen fallen größer aus, als sie bei geeigneter Schichtung wären.
2. Es besteht die Gefahr, dass es der ungeschichteten Stichprobe a priori an Repräsentativität mangelt.

# 4 Resümee

Die Analyse des 4KVen-Datenstandes unter stichprobentheoretischen Gesichtspunkten begründet wegen des Mangels an Schichtung und eines hohen zu vermutenden Designeffekts bei der Schätzung der Morbiditätsveränderungsrate große Zweifel an der Eignung dieses Datenbestandes.

Vorzuziehen wäre eine Berechnung aus einer einfachen Zufallsstichprobe von Versicherten.

Sollte die Morbiditätsveränderungsrate dennoch aus dem 4KVen-Datenstand ermittelt werden, wäre das Institut des Bewertungsausschusses vor dem Hintergrund der voran stehenden Erörterungen gut beraten, wenn es

1. die Kalibrierung des Regressionsmodells unter Berücksichtigung der Klumpenstruktur durchführt (z. B mit Hilfe der Prozedur Surveyreg von SAS), weil die Schätzungen der Beta-Gewichte dann mit korrekten Konfidenzintervallen versehen werden können.
2. die Standardabweichung der Veränderungsrate durch Anwendung von Resampling-Techniken wie JRR (Jackknife replication), BRR (Method of balanced repeated replication) oder Bootstrapverfahren schätzt (vgl. Kish/Frankel, 1974), um die Punktschätzung der Veränderungsrate mit einer Fehlerbreite absichern zu können.

# Literatur

Cochran, W. G. (1972): Stichprobenverfahren. De Gruyter, Berlin New York

Hinderer K. (1972): Grundbegriffe der Wahrscheinlichkeitstheorie. Springer, Berlin Heidelberg New York

Kish L. / Frankel M. R. (1974): Inference from Complex Samples. Journal of the Royal Statistical Society, Series B, Vol. 36,. 1-37

Kish L.: Methods for Designeffects (1995): Journal of Official Statistics 11, 55-77

Holt D. / Scott A. J. (1981): Regression Analysis Using Survey Data. The Statistician, Vol. 30, No. 3, 169-178

Schäfer T. (2004): Stichprobenverfahren. In: Voß W, et al. (Hrsg.): Taschenbuch der Statistik. 2., verbesserte Auflage. Fachbuchverlag Leipzig (im Carl Hanser Verlag), Leipzig

Sukhatme, P. V./ Sukhatme, B. V. (1970): Sampling Theory of Surveys with Applications. Ames, Iowa