



Bundesversicherungsamt

## **Vierte Bestimmung**

**des Bundesversicherungsamtes**

**nach § 273 Absatz 2 Satz 5 SGB V**

**zur kassenübergreifenden Auffälligkeitsprüfung**

**im Rahmen der Sicherung der Datengrundlagen für den  
Risikostrukturausgleich (RSA)**

**in der Fassung vom 23.10.2017**

# Inhalt

<b>VORBEMERKUNG</b> .....	<b>4</b>
<b>1 INHALT UND ANWENDUNGSBEREICH</b> .....	<b>5</b>
<b>2 RAHMENBEDINGUNGEN DER AUFFÄLLIGKEITSPRÜFUNG</b> .....	<b>6</b>
2.1 Prüfgegenstand .....	6
2.2 Vergleichskenngrößen und -zeitpunkte.....	7
2.3 Versicherten-Klassifikationsmodell als Berechnungsgrundlage .....	8
<b>3 ANALYSESCHWERPUNKT UND PRÜFAUFBAU</b> .....	<b>9</b>
<b>4 GRUPPIERUNG</b> .....	<b>10</b>
<b>5 REGRESSIONSANALYSE</b> .....	<b>12</b>
5.1 Vorbemerkungen .....	12
5.2 Kurzerklärung des Modells .....	12
5.3 Regressionsgleichung.....	13
5.4 Analysegröße.....	15
<b>6 SCHWELLENWERTE</b> .....	<b>17</b>
<b>7 MITTEILUNGEN</b> .....	<b>19</b>
7.1 Dokumentation.....	19
7.2 Bekanntgabe der Ergebnisse.....	19
<b>8 PRÜFKONSEQUENZEN, GESONDERTE BESTIMMUNGEN</b> .....	<b>20</b>
<b>9 ANHANG: AUSFÜHRLICHE DARSTELLUNG DES VERWENDETEN REGRESSIONSMODELLS</b> .....	<b>21</b>
9.1 Generalized Linear Models .....	21
9.2 Poisson-Verteilung.....	22
9.3 Linearer Prädiktor und Link-Funktion .....	23
9.4 Hauptmodell.....	23
9.5 Interpretation der Regressionskoeffizienten.....	26

Gemäß § 273 Absatz 2 Satz 5  
bestimmt das Bundesversicherungsamt

nach Zustimmung des GKV-Spitzenverbandes am xx.xx.2017

das Nähere

über die Prüfung zur Feststellung einer Auffälligkeit  
als kassenübergreifende Vergleichsanalyse

wie folgt:

## Vorbemerkung

Nach § 273 SGB V überprüft das Bundesversicherungsamt (BVA) die Zulässigkeit der Erhebung und der Verwendung von Daten im Rahmen des Risikostrukturausgleichs hinsichtlich der Vorgaben des § 268 Absatz 3 Satz 1, 2 und 14 SGB V. Gegenstand dieser Überprüfung ist insbesondere die Zulässigkeit der Meldung von Diagnosedaten entsprechend den Vorschriften zur Übermittlung und Aufbereitung von Leistungsdaten.

§ 273 SGB V sieht dazu ein dreistufiges, standardisiertes Prüfverfahren vor: In einem ersten Schritt wird eine kassenübergreifende Vergleichsanalyse durchgeführt. Werden dabei in der Datenmeldung einzelner Krankenkassen signifikante Auffälligkeiten festgestellt, kommt es bei diesen zu Einzelfallprüfungen. Erhärtet sich in einer Einzelfallprüfung der in der Auffälligkeitsprüfung aufgekommene Manipulationsverdacht, folgt im letzten Schritt die Ermittlung und Verhängung eines Korrekturbetrages.

Für Diagnosedaten aus der vertragsärztlichen Versorgung ist die Auffälligkeitsprüfung verpflichtend und wird vom BVA seit dem Berichtsjahr 2008 durchgeführt. Die Überprüfung der Diagnosedaten aus dem Krankenhaussektor ist fakultativ und erfolgte erstmals bezogen auf die gemäß § 30 Absatz 4 Satz 2 RSAV gemeldeten Daten (Korrekturmeldung) aus dem Berichtsjahr 2008.

Das Nähere zur Durchführung der Auffälligkeitsprüfung ab dem Berichtsjahr 2011 wird nachfolgend bestimmt. Gegenüber der Dritten Bestimmung zur kassenübergreifenden Auffälligkeitsprüfung in der Fassung vom 10.06.2015 ergeben sich inhaltliche Veränderungen im Hinblick auf den Zuordnungsalgorithmus stationärer Nebendiagnosen. Das Benehmen mit dem Spitzenverband Bund der Krankenkassen (GKV-Spitzenverband) gemäß § 273 Absatz 2 Satz 5 SGB V wurde hergestellt.

Ziffer 8 der Bestimmung ist lediglich deklaratorisch.

# 1 Inhalt und Anwendungsbereich

Nach § 273 Absatz 2 Satz 5 SGB V bestimmt das BVA im Benehmen mit dem Spitzenverband Bund der Krankenkassen das Nähere zur Durchführung der kassenübergreifenden Vergleichsanalyse, insbesondere einen Schwellenwert für die Feststellung einer Auffälligkeit.

Bezogen auf die obligatorische Prüfung der Diagnosedaten aus der vertragsärztlichen Versorgung nach § 273 Absatz 2 Satz 1 SGB V (SA600) und die fakultative Prüfung der Diagnosedaten aus dem Krankenhaussektor nach § 273 Absatz 2 Satz 2 SGB V (SA500), regelt diese Bestimmung einheitlich das Vorgehen zur Feststellung einer Auffälligkeit ab dem Berichtsjahr 2011.

Im Folgenden werden die Prüfgegenstände, die relevanten Vergleichsgrößen und -zeitpunkte sowie die maßgeblichen Analysegrößen in allgemeingültiger und übertragbarer Form spezifiziert. Damit entfällt die Notwendigkeit, die Bestimmung für unterschiedliche Prüfgegenstände jeweils neu zu fassen.

Die Festlegung der Schwellenwerte für eine Auffälligkeit bildet hiervon eine Ausnahme: Die Schwellenwerte werden für die stationären Hauptdiagnosen der SA500 und die kombinierten Ergebnisse der SA600 ein- und ausschließlich stationärer Nebendiagnosen der Erst- und Korrekturmeldung jeweils einzeln bestimmt. Dies geschieht in der entsprechenden Version der Anlage 1 (vgl. Abschnitt 6).

## 2 Rahmenbedingungen der Auffälligkeitsprüfung

### 2.1 Prüfgegenstand

Gegenstand der Auffälligkeitsprüfung sind jeweils die

- gemäß § 30 Absatz 4 Satz 1 RSAV gemeldeten Daten (Erstmeldung) sowie die
- gemäß § 30 Absatz 4 Satz 2 RSAV gemeldeten Daten (Korrekturmeldung)

eines Berichtsjahres nach Anwendung des BVA-Datenbereinigungskonzeptes.

Jede Prüfung erstreckt sich auf drei Prüfbereiche:

- 1) Die Hauptdiagnosen aus dem Krankenhaussektor (SA500) sowie die stationären Nebendiagnosen (SA500), die gemäß den jeweils gültigen Festlegungen nach § 31 Absatz 4 RSAV als Hauptdiagnosen gewertet werden.
- 2) Die Diagnosen aus der vertragsärztlichen Versorgung (SA600).
- 3) Die Diagnosen aus der vertragsärztlichen Versorgung (SA600) einschließlich der stationären Nebendiagnosen (SA500), die nicht in den Prüfbereich 1) fallen.

Um den Ablauf der Prüfungen zu beschleunigen, kann das BVA die Prüfungen für ein und dasselbe Berichtsjahr zusammenlegen. Dies bedeutet, dass sämtliche Berechnungen, die im Rahmen der Auffälligkeitsprüfung für ein Berichtsjahr anzustellen sind, gebündelt und die zugehörigen Schwellenwerte sowie die daraus resultierenden Prüffeststellungen parallel veröffentlicht bzw. den Krankenkassen mitgeteilt werden. Gleichwohl bleiben die Prüfung der Erstmeldung und die Prüfung der Korrekturmeldung inhaltlich eigenständig und unverbunden.

#### Hinweis:

Der Gesetzgeber sieht in § 273 SGB V eine vollständige Prüfung der Diagnosedaten vor, weswegen sowohl Erst- als auch Korrekturmeldung eines Berichtsjahres zu prüfen sind. Dies deckt sich im Übrigen mit § 39a Absatz 1 RSAV, welcher explizit ein unterschiedliches Vorgehen bei der Berechnung von (möglichen) Korrekturbeträgen in Abhängigkeit des zugrundeliegenden Meldetyps (Erst- bzw. Korrekturmeldung) vorgibt.

Die Überprüfung der Arzneimittelkennzeichen (SA400) findet gemäß des in Abschnitt 4 genannten Zuordnungsalgorithmus bei dem jeweiligen Prüfbereich zur HMG-

Zuordnung statt, so dass die Entwicklung der verordneten Arzneimittel indirekt bei der Prüfung berücksichtigt wird.

## 2.2 Vergleichskenngrößen und -zeitpunkte

Die Prüfung beruht auf dem zeitlichen Vergleich der zu prüfenden mit der vorangegangenen Meldung (im Folgenden Prüf- und Referenzmeldung genannt). Es gilt also der folgende Zusammenhang:

- Ist die Prüfmeldung eine Erstmeldung, wird die Korrekturmeldung aus dem vorangegangenen Berichtsjahr als Referenzmeldung herangezogen.
- Ist die Prüfmeldung eine Korrekturmeldung, wird die Erstmeldung desselben Berichtsjahres als Referenzmeldung herangezogen.

Sofern das BVA auf Basis der Auffälligkeitsprüfung im weiteren Prüfverlauf Verstöße i. S. d. § 273 Absatz 4 Satz 2 SGB V im Bereich der Erstmeldung feststellt, werden diese Feststellungen auf die entsprechende Korrekturmeldung übertragen und leiten automatisch eine anlassbezogene Einzelfallprüfung nach § 273 Absatz 3 Satz 2 SGB V ein. Dies ist bei simultaner Auffälligkeitsprüfung (von Erst- und Korrekturmeldung) erforderlich, da andernfalls Verstöße, die sich in der Korrekturmeldung fortsetzen, nicht zu ahnden wären: die reguläre Überprüfung der Korrekturmeldung auf statistische Auffälligkeiten bliebe in solchen Fällen wegen des Manipulations-Basiseffektes in der Referenzmeldung von vornherein ergebnislos.

### Hinweis:

Wenngleich es sich in beiden Fällen um die direkt vorangehende Meldung handelt, ist die Kontinuität von Erst- zu Korrekturmeldung eines Berichtsjahres wesentlich höher als von der Korrekturmeldung eines Berichtsjahres zur Erstmeldung des darauf folgenden Berichtsjahres. So ist grundsätzlich von einer geringen unterjährigen Veränderung der Diagnosezahlen auszugehen, wohingegen ein Berichtsjahrwechsel i. d. R. einen markanten Diagnoseanstieg mit sich bringt, wie alle bislang verfügbaren Zeitreihen belegen. Hinzu kommt, dass in dem kurzen, 10-monatigen Zeitraum zwischen der Erhebung von Erst- und Korrekturmeldung nachträgliche Änderungen der Diagnosedaten rechtlich nur in sehr begrenzten Fallkonstellationen zulässig sind. Diese Asymmetrie im Verhältnis der beiden Meldetypen führt selbst bei gleichbleibender Methodik zu strukturell stark unterschiedlichen Ergebnissen bei der Prüfung von Erst- und Korrekturmeldung eines Berichtsjahres. Diesem Umstand wird durch die Bestimmung individueller Schwellenwerte in Abhängigkeit des jeweiligen Prüfgegenstandes Rechnung getragen (vgl. Abschnitt 6).

## 2.3 Versicherten-Klassifikationsmodell als Berechnungsgrundlage

Prüf- und Referenzmeldung werden für die Prüfung auf Basis des Versicherten-Klassifikationsmodells gemäß § 31 Absatz 4 RSAV aufbereitet.

Für jedes Ausgleichsjahr existiert eine eigene Version des Klassifikationsmodells. Somit richtet sich die Wahl des Klassifikationsmodells nach dem Ausgleichsjahr, auf das sich die Prüfmeldung bezieht. Dies ist – aufgrund der prospektiven Modellausrichtung – stets das Jahr, das auf das Berichtsjahr der Prüfmeldung folgt.

Für jede Prüfmeldung ab dem Berichtsjahr 2009 dokumentiert Anlage 2 das jeweils verwendete Klassifikationsmodell mit genauen Angaben zu Ausgestaltungsdetails und technischen Spezifikationen. Seit dem Ausgleichsjahr 2012 werden stationäre Nebendiagnosen einer strikteren Validierung unterzogen, um die Manipulationsresistenz des Klassifikationsmodells zu erhöhen. Stationäre Nebendiagnosen werden seither in den meisten Fällen den Diagnosen aus der vertragsärztlichen Versorgung gleichgestellt. Dadurch werden stationäre Nebendiagnosen prinzipiell bei der Prüfung der ambulanten Aufgreifkriterien M2Q sowie den Arzneimittelkriterien mit berücksichtigt. Ausnahmen hiervon sind in den jeweils gültigen Festlegungen nach § 31 Absatz 4 RSAV dargelegt. Aufgrund der weitgehenden Gleichstellung stationärer Nebendiagnosen und Diagnosen aus der vertragsärztlichen Versorgung seit dem Ausgleichsjahr 2012 wird für die Prüfung der Datenmeldungen ab dem Berichtsjahr 2011 ein zusätzlicher Prüfbereich – SA600 einschließlich stationärer Nebendiagnosen – eingeführt. Über diese Prüfung werden auffällige Anstiege in HMG-Besetzungszahlen im Hinblick auf die zuweisungsrelevante Zuordnung der Diagnosen im Klassifikationsmodell erfasst.

### Hinweis:

Da SA500 und SA600, wie alle anderen Meldeformate, als Datengrundlage für die Zuordnung der Versicherten im Klassifikationsmodell vorgesehen sind, dienen diese – inklusive jeweiligem Zuordnungsalgorithmus für die Diagnosen - als Bezugspunkte der Analyse. Zugleich wird mit diesem Fokus den Vorgaben aus § 273 Absatz 2 Satz 4 SGB V entsprochen, welcher ausdrücklich eine Überprüfung der Daten im Hinblick auf die Klassifikation der Versicherten nach Morbidität nach § 268 Absatz 1 Satz Nr. 1 SGB V vorsieht.



### 3 Analyseschwerpunkt und Prüfaufbau

Der Analyseschwerpunkt liegt auf der Entwicklung der HMG-Besetzung (Anzahl der Versicherten mit einer HMG-Zuordnung) von der Referenz- zur Prüfmeldung auf Einzelkassenebene. Dabei wird eine separate Gruppierung der Hauptdiagnosen der SA500 (unter Berücksichtigung der Sonderfälle im Bereich stationärer Nebendiagnosen) und der SA600 jeweils ein- und ausschließlich stationärer Nebendiagnosen-vorgenommen.

Aus dem Vergleich der Gruppierungsergebnisse für Prüf- und Referenzmeldung lassen sich die für die eigentliche Auffälligkeitsprüfung relevanten Kennzahlen erzeugen. Dies geschieht auf Basis einer Regression, die für die drei Prüfbereiche und sämtliche HMGs durchgeführt wird (vgl. Abschnitt 5). Die angewandte Prüfmethodik ist dabei für die drei Prüfbereiche und für sämtliche HMGs bis zur Anwendung des Schwellenwertes identisch. Unter Anwendung der unter Abschnitt 6 definierten Schwellenwerte lassen sich auf dieser Grundlage im letzten Schritt die auffälligen Krankenkassen identifizieren. Dabei wird der Schwellenwert bei der Prüfung der SA600 ein- und ausschließlich stationärer Nebendiagnosen auf die kombinierten Ergebnisse angewandt.

#### Hinweis:

Die insgesamt mehr als 130 HMGs stellen jene Zuordnungsebene im Klassifikationsmodell dar, auf deren Basis die morbiditätsorientierten Zuschläge (HMG-Zuschläge) je Versicherten vergeben werden. Es ist daher davon auszugehen, dass sich manipulationsbedingte Abweichungen in den zu überprüfenden Daten auf dieser Ebene am ausgeprägtesten manifestieren.

Gleichzeitig bedeutet die Aggregation der Diagnose- und auch der Arzneimitteldaten in der Gruppierung entsprechend dem für das Ausgleichsjahr gültigen Zuordnungsalgorithmus eine Verdichtung der aus den rohen Morbiditätsdaten zu gewinnenden Informationen. Dabei stellen die HMGs als Analyseebene insbesondere sicher, dass die Auffälligkeitsprüfung als Kriterium auch die unterschiedlichen Schweregrade der Diagnosen berücksichtigt, vgl. § 273 Absatz 2 Satz 4 SGB V.

## 4 Gruppierung

Prüf- und Referenzmeldung werden nach den Vorgaben des unter 2.3 spezifizierten Klassifikationsmodells gruppiert. Dabei gelten folgende Abweichungen von der regulären Gruppierung:

- 1) Als Stammdaten-Datei wird – je nachdem, ob Erst- oder Korrekturmeldung zu prüfen ist – ausschließlich die zugehörige SA100 verwendet.
- 2) Die Diagnosen der SA500 und SA600 werden entsprechend der Definition der Prüfbereiche unter 2.1 jeweils unter Verwendung der SA400 gruppiert.
- 3) Die Versichertenzuordnung erfolgt auf Grundlage des kassenindividuellen Pseudonyms.
- 4) Die Zuordnung der Versicherten zu den Morbiditätsgruppen erfolgt zeitgleich.
- 5) Versicherte mit null Versichertentagen werden nicht berücksichtigt, d. h. ihnen werden zwar AGGs, aber keine HMGs zugeordnet.
- 6) Kostenerstatter und Auslandsversicherte (Zuordnung erfolgt ebenfalls über das kassenindividuelle Pseudonym und auf Basis der SA100) werden nicht berücksichtigt, d. h. ihnen werden keine HMGs zugeordnet.
- 7) Daten, die nach § 30 Absatz 4 Satz 4 RSAV fehlerhaft oder nach § 39a Absatz 3 Satz 1 RSAV nicht berücksichtigungsfähig sind, werden aus den Berechnungen ausgeschlossen. Stattdessen werden solche Daten – wenn vorhanden – über eine Liste mit den auszuschließenden Datensätzen (sog. „Diagnose-Streichliste“) entfernt oder durch eine bereinigte Datenmeldung (sog. „Austauschsatzart“) ersetzt.

Die Gruppierungen beziehen das gesamte vorhandene HMG-Spektrum mit ein. Ausgenommen sind nur solche HMGs, die aufgrund der Trennung zwischen den Hauptdiagnosen der SA500 und den vertragsärztlichen Diagnosen der SA600 ein- und ausschließlich stationärer Nebendiagnosen unbesetzt bleiben.

Da bei der Prüfung einer Korrekturmeldung die zugehörige Referenzmeldung (Erstmeldung aus demselben Berichtsjahr) i. d. R. einen unterschiedlichen Fusionsstand aufweist, wird die betroffene Erstmeldung stets auf den gleichen Fusionsstand wie die zu prüfende Korrekturmeldung gebracht.

### Hinweis:

Die Zuordnung zu den unter 6) genannten Gruppen weicht von der regulären Gruppierung ab, da eine Aggregation der Versichertentage auf GKV-Ebene nicht

stattfindet. Es kann also theoretisch bei Krankenkassenwechslern vorkommen, dass diese – anders als bei der regulären Gruppierung – den Status „Kostenerstatter“ oder „Auslandsversicherte“ nicht erhalten und in der Folge für diese Versicherte eine Zuordnung zu HMGs erfolgt. Es handelt sich hierbei jedoch in der Praxis allenfalls um eine äußerst seltene Randerscheinung ohne zahlenmäßige Relevanz.

## **5 Regressionsanalyse**

### **5.1 Vorbemerkungen**

Der Analyseschwerpunkt liegt auf der Veränderung der Besetzung in den einzelnen HMGs beim Übergang von Referenz- zu Prüfmeldung auf Ebene der Einzelkasse. Die Veränderung der Besetzungszahlen wird in Form von Relativen Risiken dargestellt. Diese und weitere Kennzahlen zur Identifikation der kassenbezogenen Auffälligkeiten werden mit einer Regressionsanalyse ermittelt.

Dieses statistische Verfahren dient dazu, auffällige Abweichungen sichtbar zu machen, die von einer zur nächsten Meldung in bestimmten HMG-Bereichen bei einzelnen Krankenkassen auftreten. Eine Regressionsanalyse zeichnet sich vor allem dadurch aus, dass sie den Einfluss mehrerer unabhängiger Einflussgrößen auf eine Zielvariable darzustellen vermag. So lässt sich mit einer multiplen Regression bspw. berücksichtigen, welchen Einfluss die Faktoren Alter und Geschlecht auf das erhöhte Auftreten einer bestimmten HMG bei einer Krankenkasse haben. Indem die GKV-weit ermittelten HMG-Besetzungszahlen standardmäßig als Referenzkategorie in die Regression mit einbezogen werden, ist außerdem der Bezug zur GKV-Entwicklung automatisch gegeben.

Die über einen Regressionsansatz zugänglichen Rechen- und Modellierungsoptionen sind für die hier angestrebte Analyse daher sehr wertvoll. Gleichzeitig machen sie das Verfahren deutlich komplexer und schwerer darstellbar. Den Abschnitten 5.3 und 5.4, die Modell und Analysegrößen definieren, werden daher eine verbale Erläuterung unter 5.2 und ein ausführlicher Anhang zur Seite gestellt, um das Verfahren nachvollziehbar und verständlich zu machen.

### **5.2 Kurzerklärung des Modells**

Die Regressionsanalyse wertet die Unterschiede in den Gruppierungsergebnissen von Prüf- und Referenzmeldung aus. Der getrennte Fokus zwischen den Hauptdiagnosen der SA500 und den vertragsärztlichen Diagnosen der SA600 ein- und ausschließlich stationärer Nebendiagnosen wird beibehalten und die Entwicklung in jeder HMG einzeln betrachtet. Für die drei Prüfbereiche werden folglich für sämtliche HMGs einzelne Regressionen durchgeführt.

Das Vorliegen einer HMG bildet die abhängige Variable in der Regression. Als unabhängige Variablen gehen die 40 Altersgruppen<sup>1</sup>, die in der Prüfmeldung vorhandenen Krankenkassen und die zugrundeliegende Meldung (Prüf- vs. Referenzmeldung) in die Regression ein. Um die GKV-weite Entwicklung zu berücksichtigen, wird die Gesamt-GKV als zusätzliche Krankenkasse modelliert. Die Gesamt-GKV wird somit zur Referenzkategorie in der Dimension „Krankenkasse“. Die Ergebnisse der Gruppierung werden für die Durchführung der Regression nach den genannten Merkmalen gruppiert. Genau genommen ist somit die HMG-Anzahl in der jeweiligen Gruppe die Zielvariable.

Ausschlaggebend für die Analyse ist ein weiterer Parameter, der in die Regression mit aufgenommen wird: ein Interaktionsterm der beiden Variablen „Krankenkasse“ und „Meldung“. Über ihn lässt sich die Wechselbeziehung dieser beiden Einflussfaktoren bestimmen. Er erlaubt also eine Aussage darüber, in welchem Maße der Einfluss des Meldewechsels auf die HMG-Entwicklung von der einzelnen Krankenkasse abhängt. Die allgemeine GKV-Entwicklung bildet den Maßstab für die Stärke des Effekts und die relativierenden Effekte von Alter und Geschlecht sowie Größe und Morbiditätsniveau der Krankenkasse sind über die restlichen Einflussfaktoren in der Regression berücksichtigt.

Als relevante Analysegröße werden einzig die Regressionskoeffizienten verwendet, die aus dem Interaktionsterm Krankenkasse \* Meldung für jede einzelne Krankenkasse resultieren. Mathematisch transformiert, ergibt sich aus den Regressionskoeffizienten das sogenannte Relative Risiko (RR). Darüber lässt sich eindeutig ablesen ( $RR < > 1$ ), ob der Zusammenhang zwischen Meldewechsel und HMG-Entwicklung bei einer Krankenkasse unter oder über dem GKV-Trend liegt. Als Kennzahlen ergänzt werden die RRs durch die zugehörigen p-Werte. Sie geben das statistische Signifikanzniveau der jeweiligen Regressionskoeffizienten an.

### 5.3 Regressionsgleichung

Die Zielvariable  $Y_i$ , also die Anzahl der Versicherten in der HMG<sub>x</sub>, wird mit einer **log-linearen Poisson-Regression** modelliert<sup>2</sup>. Bei einer Anzahl von  $r$  Einflussgrößen lautet deren allgemeine Form:

$$\log(E(Y_i)) = \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_r x_{ir}, \quad i = 1, \dots, n,$$

<sup>1</sup> Die Zuordnung zu den 40 Altersgruppen (AGGs) erfolgt für alle Versicherten mit mehr als null Versichertentagen in der jeweiligen Satzart 100.

<sup>2</sup> Für Näheres zur Poisson-Regression und deren Einordnung in die Klasse der GLM-Modelle s. die Abschnitte 9.1 und 9.2 im Anhang.

wobei  $E(Y_i)$  für den Erwartungswert der Zielvariablen,

$x_i$  bis  $x_r$  für die unabhängigen Variablen (bzw. deren Realisierungen  $x_{i1}$ - $x_{ir}$ ) und

$\beta_1$ - $\beta_r$  für die dazugehörigen Regressionskoeffizienten steht<sup>3</sup>.

Mit den folgenden Spezifikationen für die rechte Seite der Regressionsgleichung ist das Modell für den hier vorgesehenen Zweck vollständig definiert<sup>4</sup>:

1) Die unabhängigen Variablen werden bestimmt. Sie sind im hier betrachteten Fall sämtlich kategorial, werden also durch 0/1-kodierte Dummy-Variablen abgebildet, genau durch:

- die unabhängige Variable AGG „A“, mit 40 Ausprägungen und 40-1 Dummy-Variablen,
- die unabhängige Variable Krankenkasse „K“, mit  $m+1$  Ausprägungen<sup>5</sup> und  $m$  Dummy-Variablen (wobei  $m$  der Anzahl der Krankenkassen in der Prüfmeldung entspricht) und
- die unabhängige Variable Meldung „M“ mit 2 Ausprägungen und einer Dummy-Variablen.

2) Der sog. Offset wird ergänzt, der bei aggregierten Daten den Populationsbezug herstellt, hier also:

$\beta_G \log(G_i)$ , wobei  $G_i$  die Gesamtversichertenanzahl je Krankenkasse, AGG und Meldung darstellt, mit einem Regressionskoeffizienten  $\beta_G$  von 1.

3) Der Interaktionsterm  $I$  wird als zusätzliche unabhängige Variable ergänzt, hier also:

die unabhängige Variable  $I = M * K$ , mit ebenfalls  $m+1$  Ausprägungen und  $m$  Dummy-Variablen.

---

<sup>3</sup> Vgl. hierzu die Formel (1), Abschnitt 9.3 im Anhang. Letztere ist mit dieser identisch bis auf den Achsenabschnitt  $\beta_0$ . Auf diesen kann hier und im Weiteren verzichtet werden, da das verwendete Modell keinen Achsenabschnitt vorsieht.

<sup>4</sup> Eine Abbildung der vollständigen Regressionsgleichung findet sich im Anhang, Formel (2) im Abschnitt 9.4, dort sind auch die einzelnen Regressionsparameter näher erläutert.

<sup>5</sup> Die Anzahl der Krankenkassen beträgt  $m + 1$ , weil die Gesamt-GKV als zusätzliche Krankenkasse mit in die Regression eingeht, vgl. Abschnitt 9.4 im Anhang.

### Hinweis:

Die Poisson-Regression ist das wissenschaftlich gebräuchlichste und gängigste Modell bei der Analyse von Zähldaten (nicht-negative, diskrete Zielgröße). Dies gilt insbesondere bei gruppierten Zähldaten, wie im hier betrachteten Fall. Es ist darauf hinzuweisen, dass das Poisson-Modell hier als Standardverfahren angewandt wird. Damit geht natürlicherweise einher, dass die Annahme einer poisson-verteilten Zufallsvariable nicht in allen betrachteten Fällen exakt zutreffen kann. Dies ist jedoch unerheblich, solange die Aussagekraft der Ergebnisse hiervon nicht in nennenswertem Maße beeinträchtigt wird. Das BVA hat sich daher durch zahlreiche Voranalysen – insbesondere durch die Prüfung alternativer Verteilungsannahmen und eine Untersuchung der Randbereiche (gering besetzte HMGs) – vergewissert, dass die generalisierte Annahme einer poisson-verteilten Zufallsvariable methodisch unbedenklich ist. Das theoretisch größte Problem einer in den Verteilungsannahmen unpräzisen Modellspezifikation sind unterschätzte Standardfehler. Es gibt jedoch keine Anhaltspunkte, dass es im hier betrachteten Kontext zu systematischen Verzerrungen dieser Art kommt. Als zusätzliche vorbeugende Maßnahme wird von vornherein ein strengeres Signifikanzniveau als generell üblich bei der Bewertung der Regressionsschätzer angesetzt (vgl. Abschnitt 6), um die Validität der Ergebnisse sicherzustellen.

## 5.4 Analysegröße

Der relevante Interaktionsterm  $I$  bildet die multiplikative Verknüpfung der beiden Variablen Meldung und Krankenkasse ab, also

$$M*K1 \text{ bis } M*Km, \text{ bzw. } I1 \text{ bis } Im.$$

Dies führt bei  $m$  Krankenkassen zu  $m$  Regressionskoeffizienten, also  $\beta I_1 - \beta I_m$ .

Werden die  $\beta I$ -Parameter exponiert, berechnet man also  $\exp\{\beta I_1\}$  bis  $\exp\{\beta I_m\}$ , ergibt sich das relative Risiko<sup>6</sup>:

$$RR(I_1) - RR(I_m), \text{ bzw. } RR(I_k), \text{ mit } k = 1, \dots, m. \quad (1)$$

---

<sup>6</sup> Für Näheres zum Relativen Risiko und der Interpretation der Regressionskoeffizienten s. Abschnitt 9.5 im Anhang.

Für jeden der  $\beta I$ -Parameter geben die zugehörigen p-Werte das jeweilige Signifikanzniveau an,  $\beta I_1 - \beta I_m$  werden also ergänzt durch:

$$p(\beta I_1) - p(\beta I_m), \text{ bzw. } p(\beta I_k) \text{ mit } k = 1, \dots, m. \quad (2)$$

Die Kennzahlen aus (1) und (2) werden – bezogen auf die drei Prüfbereiche – für sämtliche HMGs erzeugt (für jede  $HMG_x$  wird also – je Satzart – eine separate Regression durchgeführt).

Hinweis:

Die  $RR_s$  geben an, wie sich die Besetzungszahlen in der jeweiligen HMG bei der betreffenden Krankenkasse gemessen am GKV-Schnitt beim Übergang von Vergleichs- zu Prüfmeldung entwickelt haben. Diese Entwicklung kann überproportional ( $RR \{I_k\} > 1$ ) oder unterproportional ( $RR \{I_k\} < 1$ ) sein. Die  $p(\beta I_k)_s$  liefern hierzu das Signifikanzniveau und geben somit an, ob es sich um eine statistisch signifikante Abweichung handelt.



## 6 Schwellenwerte

Die im Rahmen der Auffälligkeitsprüfung berechneten Analysegrößen geben für sich alleine noch keine Auskunft darüber, wann eine Datenmeldung als auffällig eingestuft werden muss. Es bleiben unvermeidlich Grenzwerte bzw. Toleranzbereiche zu definieren, die den Normbereich der Untersuchung bilden. Liegen die errechneten Werte der Analysegrößen außerhalb dieser Toleranzbereiche, so gelten die Daten als auffällig und eine weitergehende Einzelfallprüfung wird eingeleitet.

Grundsätzlich gilt als notwendige Bedingung für eine Auffälligkeit, dass

$$1) \quad RR\{I_k\}_{HMGx} > 1,$$

bei einem kritischen Wert für den korrespondierenden  $\alpha$ -Fehler von 0,01, wenn also zusätzlich gilt, dass:

$$2) \quad p\{\beta I_k\}_{HMGx} < 0,01.$$

Dies sind notwendige, aber keine hinreichenden Bedingungen für eine Auffälligkeit. Angesichts der hohen Anzahl an untersuchten HMGs und der isolierten Prüfung von Hauptdiagnosen der SA500 und vertragsärztlichen Diagnosen der SA600 ein- und ausschließlich stationärer Nebendiagnosen ist es möglich, dass die Gesamtanzahl der Krankenkassen, die die notwendige Bedingung für eine Auffälligkeit erfüllen, eine kritische Größe überschreitet. Dies kann insbesondere bei der Prüfung einer Erstmeldung der Fall sein, bei der gegen die Korrekturmeldung aus dem zurückliegenden Berichtsjahr geprüft wird und es erfahrungsgemäß zu großen Abweichungen zwischen Prüf- und Referenzmeldung kommt.

In diesem Fall ist es möglich, den unter 1) genannten Schwellenwert so zu verschärfen, dass für die Prüfung der stationären Hauptdiagnosen der SA500 und die Prüfungen der SA600 ein- und ausschließlich stationärer Nebendiagnosen die Anzahl auffälliger Krankenkassen höchstens den Anteil von 15% der in der jeweiligen Meldung vertretenen Krankenkasse erreicht. Aufgrund von Überschneidungen in den geprüften Datengrundlagen bei der reinen SA600 und jener einschließlich stationärer Nebendiagnosen werden die Relativen Risiken bei der Anwendung des Schwellenwertes einer gemeinsamen Bewertung unterzogen. Dies birgt den Vorteil, dass die in der Gesamtbetrachtung beider Bereiche auffälligsten Krankenkassen, d. h. diejenigen mit den höchsten Relativen Risiken selektiert werden. Die Festsetzung der Schwellenwerte geschieht im Einvernehmen zwischen BVA und GKV-Spitzenverband auf Grundlage der in Anlage 1 enthaltenen Werte. Anlage 1 wird für jede zu prüfende Datenmeldung neu erstellt. Die jeweilige Version gibt für jede Meldung, jeden Prüfbereich (SA500 Hauptdiagnosen, SA600 ein- und ausschließlich stationärer

Nebendiagnosen) und über alle HMGs die ermittelten RRs und die zugehörigen p-Werte je Krankenkasse auf pseudonymisiert Basis an.

Die auf diese Weise für jede Prüfung zwischen BVA und GKV-Spitzenverband einvernehmlich und bindend festgelegten Schwellenwerte für 1) und 2) werden in den entsprechend gekennzeichneten Feldern der Anlage 1 veröffentlicht.

## **7 Mitteilungen**

### **7.1 Dokumentation**

Zur Dokumentation der Prüfungen erhält der GKV-Spitzenverband vom BVA die folgenden Mitteilungen, die zur Weiterleitung an die Krankenkassen vorgesehen sind:

Grundlegend für die Dokumentation des Verfahrens sind die Anlagen 1 und 2, die für jede zu prüfende Datenmeldung neu erstellt bzw. ergänzt werden. Gemeinsam werden sie dem GKV-Spitzenverband mit einem Anschreiben übermittelt, in dem das BVA die Aufnahme neuer Prüfhandlungen bekannt gibt. Anlage 1 erhält zusätzlich zu den in Abschnitt 6 genannten Werten die in Abschnitt 9.4 erwähnten Gütekriterien für sämtliche HMG-bezogenen Regressionsrechnungen.

Zusätzlich stellt das BVA dem GKV-Spitzenverband für jedes Berichtsjahr im Nachgang zur Einleitung der Prüfhandlungen einen fiktiven Testdatensatz zur Verfügung, der es den Krankenkassen ermöglichen soll, das Verfahren auch ohne Kenntnis der Gesamt-GKV-Daten rechnerisch nachzuvollziehen.

Nach Abschluss der Auffälligkeitsprüfung erhält der GKV-Spitzenverband zudem einen abschließenden Prüfbericht mit einer prüfbereichbezogenen, anonymisierten Zusammenstellung der Ergebnisse der Auffälligkeitsprüfung.

### **7.2 Bekanntgabe der Ergebnisse**

Die Bekanntgabe der Prüfergebnisse erfolgt kassenindividuell:

Anlage 1 wird verwendet, um jeder Krankenkasse die für sie berechneten Kennzahlen (RR, p-Wert) je HMG und Prüfbereich in tabellarischer Form bekannt gegeben. Darin sind zudem die Relativen Risiken all jener Krankenkassen farblich gekennzeichnet, bei denen nach der unter 6 beschriebenen Vorgehensweise bei der Bewertung der Regressionsergebnisse eine Auffälligkeit vorliegt. Ergänzend wird jeder Krankenkasse in einem Schreiben ihr Pseudonym mitgeteilt sowie Auskunft darüber gegeben, ob bei ihr eine statistische Auffälligkeit vorliegt. Die Krankenkassen erhalten die genannten Informationen über den beim BVA eingerichteten FTP-Server.

## **8 Prüfkonsequenzen, gesonderte Bestimmungen**

Die Feststellung einer Auffälligkeit nach Maßgabe des Abschnitts 6 dieser Bestimmung in Verbindung mit § 273 Absatz 2 SGB V hat eine Einzelfallprüfung zur Folge (§ 273 Absatz 3 Satz 1 SGB V). Die Mitteilung über festgestellte Auffälligkeiten nach Abschnitt 7.2 dieser Bestimmung leitet diese Einzelfallprüfung automatisch ein. Das Ergebnis der Einzelfallprüfung wird den betroffenen Krankenkassen nach Abschluss der Prüfhandlungen separat bekannt gegeben (§ 273 Absatz 4 Satz 1, Absatz 5 Satz 1 Halbsatz 1 SGB V).

**Bonn, den 23.10.2017**

## 9 Anhang: Ausführliche Darstellung des verwendeten Regressionsmodells

Im Folgenden wird die Ausgestaltung des unter Abschnitt 5 aufgeführten Regressionsmodells schrittweise erläutert. Ziel ist es, den Modellaufbau nachvollziehbar und transparent zu machen. Es geht hierbei ausdrücklich nicht um eine formal-mathematische Darstellung, da der Erklärungsgehalt von komplizierten Formeln für gering gehalten wird und entsprechende Ausführungen in statistischen Standardwerken und wissenschaftlichen Artikeln in großer Zahl zu finden sind.

### 9.1 Generalized Linear Models

Im Rahmen der Analyse wird als Zielvariable das Vorliegen bzw. Nicht-Vorliegen einer HMG je Versicherten modelliert. Da im Rahmen von § 273 SGB V die Betrachtung auf Kassenebene ausschlaggebend ist, werden die Ereignisse je Krankenkasse gezählt: Es geht folglich um die Auswertung einer Zählvariablen. Bei einer solchen Analyse, die das Vorkommen eines Ereignisses gemessen an einer bestimmten Population untersucht, sind einfache (multiple) Regressionsverfahren nicht mehr anwendbar<sup>7</sup>. Stattdessen kommen in solchen Fällen Modelle vom Typ der „Generalized Linear Models“ zum Einsatz, sogenannte GLM-Modelle.

GLM-Modelle sind eine Verallgemeinerung der einfachen linearen Modelle. Von diesen ausgehend umfasst die Verallgemeinerung drei Komponenten<sup>8</sup>:

- 1.) Die Verteilung der Zielvariablen  $Y_i$  gehört einer einfachen Exponentialfamilie an.
- 2.) Die erklärenden Variablen gehen als Linearkombination ins Modell ein:

$$\eta = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_r x_{ir}, \quad i = 1, \dots, n.$$

- 3.) Der Erwartungswert von  $Y_i$  ist durch die Linkfunktion  $g$  mit  $\eta$  verknüpft:

$$g(E(Y_i)) = \eta_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_r x_{ir}, \quad i = 1, \dots, n.$$

---

<sup>7</sup> Bei einer Zählvariablen sind viele Voraussetzungen, die an die Anwendung einer einfachen linearen Regression gekoppelt sind, nicht mehr erfüllt: So kann bei einer Zielvariable, die nur ganzzahlige Werte  $\geq 0$  annehmen darf, nicht grundsätzlich von einer Normalverteilung ausgegangen werden. Auch ist die Varianz der Werte, die die Zielvariable  $Y$  annehmen kann, nicht konstant, sondern i.d.R. ansteigend mit dem Mittelwert.

<sup>8</sup> Die Darstellung ist entlehnt aus: Müller, Marianne: „Angewandte statistische Regression“, Zürcher Hochschule Winterthur 2006, online abrufbar unter: [http://stat.ethz.ch/education/semesters/WS\\_2006\\_07/asr/Stat.pdf](http://stat.ethz.ch/education/semesters/WS_2006_07/asr/Stat.pdf).

In erster Linie lockern GLM-Modelle damit im Vergleich zu einfachen linearen Modellen die Annahmen hinsichtlich der Verteilung der Zielvariablen. So gehören zur exponentiellen Familie eine ganze Reihe von Verteilungen, bspw. die Binomial-, Poisson- oder Gamma-Verteilung. Die Flexibilisierung erfolgt in einem einheitlichen methodischen Rahmen. Eine wesentliche Komponente dabei ist die Link-Funktion: Da GLM-Modelle – genauso wie einfache lineare Regressionen – linear in ihren Koeffizienten sind, bedarf es der Link-Funktion, um den Erwartungswert  $E(Y_i)$  in Werte des linearen Prädiktors zu verwandeln.

## 9.2 Poisson-Verteilung

Für den hier vorliegenden Fall wird davon ausgegangen, dass die Zielvariable  $Y_i$  (Häufigkeit des Vorliegens einer HMG) poisson-verteilt ist. Die Poisson-Verteilung erlaubt eine Annahme darüber, wie oft voneinander unabhängige Ereignisse – bei einer gegebenen Eintrittswahrscheinlichkeit  $p$  – in einer bestimmten Zeit oder in einem bestimmten Volumen auftreten. Dabei lässt sich die gesuchte Anzahl direkt vom erwarteten Mittelwert  $\mu$  der Ereignisse ableiten. Der Parameter  $\mu$  charakterisiert folglich die Poisson-Verteilung. Er ist gleich der mittleren Anzahl der gezählten Ereignisse, die erwartet wird, wenn die Zählung  $n$  Male wiederholt wird. Wenn die Zielvariable  $Y_i$  also poisson-verteilt ist mit dem Parameter  $\mu$ , dann ist ihr Erwartungswert  $E(Y_i) = \mu_i$ , mit der Restriktion  $\mu_i > 0$ .

Die Poisson-Verteilung ist also für diskrete Werte größer gleich null definiert, weswegen sie sich für die Analyse von Zählgrößen anbietet. Dies gilt insbesondere, wenn es sich um seltene Ereignisse mit einer geringen Eintrittswahrscheinlichkeit  $p$  handelt und die Population, auf die sich die Wahrscheinlichkeit bezieht, sehr groß ist (mit vierstelligem  $n$  oder größer)<sup>9</sup>. Folglich spricht man bei der Poisson-Verteilung auch häufig von der „Verteilung der seltenen Ereignisse“. Gerade bei medizinischen Inzidenzraten, die hier betrachtet werden, lässt sie i. d. R. eine gute Annäherung an die realen Daten zu.

Eine Besonderheit der Poisson-Verteilung ist die Eigenschaft, dass nicht nur der Erwartungswert, sondern auch die Varianz gleich dem Mittelwert  $\mu$  ist. Eine weitere Eigenschaft besteht darin, dass die Summe von unabhängigen poisson-verteilten

---

<sup>9</sup> Die Poisson-Verteilung kann auch als Grenzfall der Binomialverteilung aufgefasst werden, nämlich für  $n \rightarrow \infty$ . Dann konvergiert die Dichte der Binomialverteilung gegen die der Poisson-Verteilung. Dies trifft auf die zugrundeliegenden Daten praktisch zu. Entsprechend haben empirische Tests gezeigt, dass Binomial- und Poisson-Regression im hier vorgefundenen Kontext (nahezu) identische Ergebnisse produzieren.

Ereignissen ebenfalls poisson-verteilt ist. Folglich kann man Individualdaten und gruppierte Daten gleichermaßen mit einer Poisson-Regression untersuchen.

### 9.3 Linearer Prädiktor und Link-Funktion

Sobald die Verteilung von  $Y_i$  festgelegt ist, kann die eigentliche Regressionsgleichung aufgestellt werden. Als Ausgangspunkt besitzt auch die Poisson-Regression den – unter 9.1 vorgestellten – „Linearen Prädiktor“:

$$\eta_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_r x_{ir}, i = 1, \dots, n.$$

In der Poisson-Regression wird als Link-Funktion – wegen der Restriktion  $\mu_i > 0$  – der Logarithmus des Erwartungswertes verwendet, denn er macht aus den positiven Erwartungswerten transformierte Werte ohne Begrenzung. Es gilt also:

$$g(E(Y_i)) = \eta = \log(E(Y_i)),$$

bzw.:

$$\log(E(Y_i)) = \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_r x_{ir}, i = 1, \dots, n. \quad (1)$$

Die sog. **log-lineare Poisson-Regression** kombiniert somit die logarithmische Link-Funktion mit der Annahme der Poisson-Verteilung für die Zielgröße.

### 9.4 Hauptmodell

Um die spezifische Regressionsgleichung aufstellen zu können, bedarf es ausgehend von Gleichung (1) noch dreier weiterer Schritte<sup>10</sup>:

- 1) Bestimmung der unabhängigen Variablen, bzw. Kovariablen.
- 2) Bestimmung des sog. Offsets.
- 3) Bestimmung des Interaktionsterms aus den Kovariablen „Meldung“ und „Krankenkasse“.

---

<sup>10</sup> Es folgt eine ausführliche Schilderung der unter 5.3 angegebenen Bestandteile der verwendeten Regressionsgleichung.

## Zu 1) Bestimmung der unabhängigen Variablen

Um den Einfluss auf die Zielvariable HMG-Häufigkeit zu messen, stehen drei Merkmale zur Verfügung: Im Zentrum der Analyse stehen die Einzelkassen, die HMG-Besetzungszahlen werden also nach den in der Prüfmeldung vorhandenen Krankenkassen differenziert. Bei einer Anzahl von  $m$  Krankenkassen geht hierbei die GKV als  $m + 1$ te Krankenkasse mit in die Betrachtung ein und bildet die relevante Referenzgruppe, mit der verglichen wird. Weiter gehen die Merkmale Alter und Geschlecht in Form der 40 AGGs in die Regression mit ein. Das letzte und entscheidende Merkmal, das als unabhängige Variable ins Modell einfließt, ist die Meldung (Referenz- vs. Prüfmeldung).

Da es sich bei allen Merkmalen um kategoriale Variablen handelt, gehen sie als 0/1-kodierte Dummy-Variablen in das Modell ein. Programmtechnisch erfolgt diese Umwandlung i.d.R. automatisch. Für das Merkmal „Meldung“, das ohnehin als dichotome Variable vorliegt, erübrigt sich dieser Schritt.

Somit wird der lineare Prädiktor gebildet aus:

- der unabhängigen Variablen AGG „ $A$ “, mit 40 Ausprägungen und 40-1 Dummy-Variablen,
- der unabhängigen Variablen Krankenkasse „ $K$ “, mit  $m+1$  Ausprägungen und  $m$  Dummy-Variablen, und
- der unabhängigen Variable Meldung „ $M$ “ mit 2 Ausprägungen und einer Dummy-Variablen.

## Zu 2) Der offset

Die Bestimmung der unabhängigen Variablen erleichtert die Handhabung des Regressionsdatensatzes: Die Gruppierungen, die für Prüf- und Referenzmeldung getrennt und auf Individualdaten-Basis vorliegen, werden zusammengespielt und nach diesen Merkmalen gruppiert. Angesichts der  $m+1$  Krankenkassen, der 40 AGGs und der zwei einbezogenen Meldungen bedeutet dies eine Verschlinkung des Regressionsdatensatzes auf  $(m+1) \times 40 \times 2$  Beobachtungen, ohne dass ein Informationsverlust entsteht.

Darüber ergeben sich jedoch HMG-Raten, denn die Anzahl der Ereignisse pro Krankenkasse, Altersgruppe und Meldung  $Y_i$  lässt sich nur im Verhältnis zur Gesamtversichertenanzahl  $G_i$  in der jeweiligen Gruppe sinnvoll interpretieren. Da dieses



Verhältnis – also die erwartete Anzahl Erkrankter zur Gesamtgruppe – exakt proportional ist, wäre es daher auch als direkte Zielgröße geeignet. Dies ist jedoch ausgeschlossen, da  $Y_i/G_i$  nicht mehr poisson-verteilt ist. Deshalb muss die Regression den jeweiligen Populationsbezug durch ein spezielles Argument vorsehen. Dies wird – programmtechnisch – als offset bezeichnet. Hierzu gehen die Werte für  $G_i$  logarithmiert in den eigentlichen Regressionsterm mit ein. Der lineare Prädiktor erhält also das zusätzliche Argument:

$$\beta_G \log(G_i),$$

mit dem Koeffizienten  $\beta_G = 1$ .

### Zu 3) Interaktionsterm

Um die eigentliche Analysegröße für das Erkennen einer Auffälligkeit aus der Regression gewinnen zu können, muss dem linearen Prädiktor noch ein weiteres Element hinzugefügt werden: Ein Interaktionsterm der beiden Kovariablen Krankenkasse und Meldung, der die Wechselwirkungen zwischen diesen beiden Einflussgrößen darzustellen vermag. Der relevante Interaktionsterm  $I$  lässt sich durch eine multiplikative Verknüpfung der beiden betreffenden Variablen ins Modell integrieren, also  $I = M * K$ .

Dabei gehen die restlichen Kovariablen  $(A, M, K)$  unverändert ins Modell mit ein. Inhaltlich stellt der Einbezug der AGGs sicher, dass um die möglicherweise verzerrenden Alters- und Geschlechtseinflüsse kontrolliert wird. Durch den Einbezug des Parameters „Krankenkasse“ als erklärende Variable wird hingegen um die verschiedenen Morbiditätsniveaus der Krankenkassen kontrolliert. Die Variable Meldung kontrolliert schließlich um den Morbiditäts-Basiseffekt, also um ein ansteigendes HMG-Grundniveau von der Referenz- zur Prüfmeldung.

Damit ergibt sich für das Hauptmodell die Regressionsgleichung:

$$\log(E(Y_i)) = \beta_G \log(G_i) + A'_i + K'_i + M'_i + I'_i, \dots, i = n, \quad (2)$$

wobei  $E(Y_i)$  für den Erwartungswert der Zielvariablen,

$\beta_G \log(G_i)$  für den Offset,

$A'_i$  verkürzt für  $(A1)_i \beta_{A1} + \dots + (A39)_i \beta_{A39}$ ,

$K'_i$  verkürzt für  $(K1)_i \beta_{K1} + \dots + (Km)_i \beta_{Km}$ ,

$M'_i$  verkürzt für  $M_i\beta_M$  und

$I'_i$  verkürzt für  $(M * K1)_i\beta_{I1} + \dots + (M * Km)_i\beta_{Im}$  steht.

Die Berechnungen werden vom BVA mit der Statistik-Software SAS durchgeführt. Für das hier spezifizierte Modell gilt wie für alle GLM-Regressionen, dass sich die Regressionskoeffizienten („ $\beta$ -Parameter“) nicht mehr mittels einer Kleinst-Quadrat-Methode exakt bestimmen lassen. Stattdessen müssen sie mit der Maximum-Likelihood-Methode in einem iterativen Prozess geschätzt werden<sup>11</sup>. Auch alle Tests und Vertrauensintervalle sind damit approximativ. Ähnlich wie auch in der Linearen Regression lässt sich die Güte der Schätzung über eine Residualanalyse bestimmen. In den gängigen Statistik-Programmen werden zu diesem Zweck i.d.R. die Pearson-Residuen und die Devianz-Residuen ausgegeben. Da bei dem hier vorliegenden Untersuchungsgegenstand nicht die Vorhersagegüte des Modells im Vordergrund steht, sondern in erster Linie der Effekt bestimmter Einflussgrößen auf die Zielvariable ermittelt werden soll, wird auf eingehende Beschreibungen zu den von SAS ausgegebenen Gütekriterien an dieser Stelle verzichtet. Zu Dokumentationszwecken ist jedoch der Quotient aus Devianz und Anzahl der Freiheitsgrade als dasjenige Kriterium, das am häufigsten zur Überprüfung der Modellgüte herangezogen wird, in der Anlage 1 je Satzart für jede HMG-bezogene Regression ausgegeben. Es sei allerdings darauf verwiesen, dass die Aussagekraft dieser Werte als gering eingeschätzt werden kann, da sie nur im Vergleich verschiedener Modellvarianten sinnvoll interpretierbar sind, nicht jedoch im Vergleich unterschiedlicher Regressionen (mit unterschiedlichen zugrundeliegenden Daten), wie im hier betrachteten Fall.

## 9.5 Interpretation der Regressionskoeffizienten

Um nun zu einer Interpretation der Regressionskoeffizienten zu gelangen, bietet es sich an, die vorangegangene Gleichung in einer Schreibweise wiederzugeben, in der nach  $E(Y_i)$  aufgelöst, bzw. beide Seiten exponiert sind, also:

$$E(Y_i) = \exp\langle\beta_{A1}\rangle^{A1} \cdot \dots \cdot \exp\langle\beta_{Im}\rangle^{Im}. \quad (3)$$

---

<sup>11</sup> Bei einer Maximum-Likelihood-Methode ist nicht grundsätzlich garantiert, dass der zugrundeliegende Algorithmus konvergiert. Der Rechenprozess im hier verwendeten Modell umfasst unabhängig davon maximal 50 Iterationen. Dies entspricht der Default-Einstellung bei SAS.

Es wird deutlich, dass die additiven Terme im Bereich des linearen Prädiktors durch Exponieren in multiplikative Effekte umgewandelt werden. D.h. ändert man z.B.  $A_1$  um eine Einheit, bewirkt dies eine Multiplikation des Erwartungswertes mit dem Faktor  $\exp\{\beta_{A_1}\}$ . Generell bedeutet  $\exp\{\beta_j\} = 1$ , dass kein Einfluss zwischen der unabhängigen und der abhängigen Variable besteht, bei  $\exp\{\beta_j\} > 1$  ist er positiv, bei  $\exp\{\beta_j\} < 1$  ist er negativ. Dabei lässt sich  $\exp\{\beta_j\}$  auch als das sog. Relative Risiko  $RR$  interpretieren. Das  $RR$  ist ein Wahrscheinlichkeitsverhältnis, das die Wahrscheinlichkeit für den Eintritt eines Ereignisses im Unterschied zwischen zwei Gruppen angibt. Würde bspw. die HMG-Anzahl nur anhand des Alters erklärt und hierfür zwei Gruppen gebildet mit der Grenze  $\geq 50$  Jahre, dann würde ein  $\exp\{\beta_A\} = RR(A) = 1,6$  bedeuten, dass die HMG- Anzahl bei über 50-Jährigen um einen Faktor von 1,6 höher ist als bei unter 50-Jährigen.

Als Analysegröße ist im Rahmen des mit Gleichung (2) abgebildeten Modells einzig der Einfluss des Interaktionsterms  $I$ , also die  $RR(I)_s$ , von Relevanz. Sie geben an, wie sich der Einfluss auf die HMG-Häufigkeit, der der Meldung (bzw. dem Übergang von Referenz- zur Prüfmeldung) zuzuschreiben ist, in Abhängigkeit unterschiedlicher Krankenkassen ändert. Ist diese Wechselbeziehung stark, das zugehörige  $RR$  bei einer Krankenkasse also (erheblich)  $> 1$ , ist der Einfluss auf die HMG-Entwicklung überproportional. Bei der entsprechenden Krankenkasse führt der Übergang von der Referenz- zur Prüfmeldung also zu einem Anstieg der Versichertenzahlen in der betrachteten HMG, der (weit) oberhalb der GKV-Entwicklung liegt (für die GKV gilt stets  $RR(I) = 1$ ). Die entsprechende p-Statistik liefert für jeden  $\beta I$ -Parameter das zugehörige Signifikanzniveau<sup>12</sup>. Beide Kennzahlen bilden gemeinsam das Maß für den Grad einer Auffälligkeit.

---

<sup>12</sup> Es handelt sich um eine SAS-generierte Wald- $\chi^2$ -Statistik unter der Annahme, dass die Maximum-Likelihood-Schätzer asymptotisch normalverteilt sind. Dies gilt gewöhnlich für große Stichprobenumfänge.