Gutachten

zur Risikopoolprüfung gemäß § 20 Abs. 1 Satz 2 RSAV

– Stichprobenmethodik und Hochrechnung –

20. Dezember 2022

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Zusammenfassung

Mit dem “Gesetz für einen fairen Kassenwettbewerb in der gesetzlichen Krankenversicherung” (GKV-FKG) wurde ein Risikopool eingeführt, dessen ordnungsgemäße Abrechnung Teil einer jährlichen Risikopoolprüfung ist. Aufgabe dieses Gutachtens ist die Ausarbeitung einer zuverlässigen Stichproben- und Hochrechnungsmethode für die anstehenden Risikopoolprüfungen. Im Risikopool sind im Startjahr 2021 nur Hochrisikoversicherte mit Ausgaben über einem Schwellenwert von 100.000€ enthalten (Grundgesamtheit).[1] Die Anzahl der Hochrisikoversicherten ist in allen Krankenkassen unterschiedlich (zwischen 1 und 14.767 Hochrisikoversicherten in 2021). Krankenkassen bekommen für jeden Hochrisikoversicherten 80% der Leistungsausgaben über dem Schwellenwert zugewiesen. Ziel der Risikopoolprüfungen ist eine verlässliche Abschätzung der Zuweisungsfehler in jeder Krankenkasse. Aufgrund der Logik im Risikopool kann zwischen dem “Abrechnungsfehler”, also jenem Eurobetrag, der sich in der Prüfung als fehlerhaft herausstellt, und dem “Zuweisungsfehler”, also jenem Betrag, den die Krankenkasse im Risikopool zu viel erhalten hat, unterschieden werden. Der Zuweisungsfehler lässt sich im Wesentlichen direkt aus dem Abrechnungsfehler herleiten und entspricht der Differenz aus den kassenspezifischen Zuweisungen vor und nach Prüfung (vgl. auch Gleichung 20). In 2021 waren 96 gesetzliche Krankenkassen aktiv. Im Interesse der Prüfgerechtigkeit soll die Genauigkeit dieser Abschätzung über alle Krankenkassen hinweg gleich gut ausfallen. Aufgrund von Prüfkapazitätsbeschränkungen ist eine Kombination aus Voll- und Stichprobenprüfung angedacht.

Nachdem eine Pilotprüfung noch nicht durchgeführt wurde, liegt zum Zeitpunkt des Gutachtens noch kein historischer Datensatz zu den Zuweisungsfehlern im Risikopool vor. Das Gutachten ist also auf Annahmen über die Zuweisungsfehler angewiesen. Um diese Annahmen möglichst plausibel zu wählen, steht ein Datensatz zu den Hochrisikofällen des Jahres 2021 zur Verfügung.[2] Im Folgenden werden die Ergebnisse des Gutachtens in vereinfachter Form zusammengefasst. Eine ausführliche Auseinandersetzung mit den Themen und den dahinterstehenden Annahmen findet sich im Hauptteil des Gutachtens. Diese Ergebnisse sollten nach einer Pilotprüfung und/oder dem ersten Prüfzyklus noch überprüft werden. Andernfalls besteht die Gefahr, die Stichproben deutlich zu groß bzw. deutlich zu klein zu wählen.

- *Übergang zwischen Voll- und Stichprobenprüfung:*
  Ein hinreichend großer Stichprobenumfang ist notwendig, um die Schätzgenauigkeit verlässlich konstant halten zu können. Der Übergang zwischen Voll- und Stichprobenprüfung sollte deswegen nicht zu niedrig gesetzt werden. Für Krankenkassen mit weniger als 50 Hochrisikoversicherten sollte von daher vorerst eine Vollprüfung in Betracht gezogen werden. Bei diesem Schwellenwert wären 2021 beispielsweise 34 Krankenkassen mit zusammen 767 Hochrisikofällen in der Vollprüfung gewesen. Mehr Details hierzu finden sich in Kapitel 4.2 (insbesondere in Tabelle 2 und Abbildung 3).

- *Stichprobenprüfung:*
  Nachdem die Grundgesamtheiten in dem neuen Prüfverfahren deutlich homogener sind als in den bestehenden Prüfungen, sind einfache Zufallsstichproben vermutlich ausrei-

[1] Der Schwellenwert wird jährlich angepasst.
[2] Eine detaillierte Beschreibung dieses Datensatzes findet sich in Kapitel 4.1.

chend. Dieser Punkt kann aber mangels Daten zu den Zuweisungsfehlern noch nicht empirisch untersucht werden. Darüber hinaus würde eine geschichtete Zufallsstichprobe bei kleinen Krankenkassen zu noch kleineren Grundgesamtheiten innerhalb der Schichten führen (Median der Hochrisikoversicherten in einer Krankenkasse liegt bei 130). Diese Punkte sprechen eher gegen geschichtete Zufallsstichproben.

- *Bestimmung des kassenspezifischen Stichprobenumfangs:*
  Um die Schätzgenauigkeit konstant zu halten, muss der Stichprobenumfang zwischen den Krankenkassen variieren. Die entscheidenden Größen für die Variation im kassenspezifischen Stichprobenumfang sind die Grundgesamtheit der Hochrisikoversicherten in einer Krankenkasse sowie der Variationskoeffizient (also das Verhältnis von Standardabweichung zu Mittelwert) der Zuweisungsfehler. Im Allgemeinen gilt, je größer der Variationskoeffizient bzw. die Grundgesamtheit, desto größer fällt die Stichprobe aus (wobei ebenfalls vorerst eine Untergrenze von mindestens 50 Hochrisikoversicherten in jeder Stichprobe vorhanden sein sollte). Die Bestimmung des Stichprobenumfangs wird in Kapitel 3 ausführlich erläutert. Die letztendlich verwendete Formel findet sich in Gleichung (14).

- *Hochrechnung:*
  Während bei der Vollprüfung die Summe der Zuweisungsfehler in der Grundgesamtheit direkt beobachtet werden kann, kann eine Stichprobenprüfung diese Summe nur auf dem Wege der Hochrechnung annäherungsweise bestimmen. Basierend auf einer hinreichend großen Stichprobe lässt sich die kassenspezifische Summe der Zuweisungsfehler verlässlich hochrechnen. Hierfür wird die monetäre Fehlerquote einer Krankenkasse mit der Anzahl der Hochrisikoversicherten in dieser Krankenkasse multipliziert (siehe Gleichung 5). Es liegt in der Natur der Sache, dass mit der Stichprobenprüfung eine Ungenauigkeit über die Summe der Zuweisungsfehler in der Grundgesamtheit einhergeht. Der kassenspezifische Stichprobenumfang ist dabei so gewählt, dass die Schätzgenauigkeit der monetären Fehlerquote über alle Krankenkassen in der Stichprobenprüfung hinweg gleich ausfällt.

- *Stichprobengüte:*
  Aus Gründen der Prüfgerechtigkeit kann es Sinn machen Stichproben, die in auffälliger Weise nicht “repräsentativ” für die Grundgesamtheit einer Krankenkasse sind, auszuschließen. Angelehnt an die bereits bestehenden Prüfverfahren kann die Stichprobengüte mittels einfachen Regressionen überprüft werden (für mehr Details siehe Kapitel 5).

1 Hintergrund und Aufgabenstellung

Mit dem “Gesetz für einen fairen Kassenwettbewerb in der gesetzlichen Krankenversicherung” (GKV-FKG) wurde unter anderem ein Risikopool eingeführt, welcher die Belastung durch Hochkostenfälle abfedern soll. Krankenkassen bekommen für jeden Leistungsfall, dessen Kosten einen bestimmten Schwellenwert überschreitet, 80% der Leistungsausgaben erstattet. Die Zuweisungen aus dem Risikopool basieren also, anders als der morbiditätsorientierte Risikostrukturausgleich, auf Ist-Ausgaben, deren ordnungsgemäße Abrechnung Teil einer jährlichen Risikopoolprüfung gemäß §20(1) Satz 2 RSAV ist. Aufgabe dieses Gutachtens, welches vom Bundesamt für Soziale Sicherung (BAS) in Auftrag gegeben wurde, ist die Ausarbeitung einer zuverlässigen Stichproben- und Hochrechnungsmethode für die anstehenden Risikopoolprüfungen.

Nachdem die Leistungsausgaben im Risikopool unmittelbar zahlungsbegründend sind, ist vom Gesetzgeber immer eine Inrechnungstellung der Zuweisungsfehler vorgesehen. Mit Inrechnungstellung gemäß §20(5) RSAV ist in diesem Gutachten eine Hochrechnung der Zuweisungsfehler gemeint. Aufgrund der Logik im Risikopool kann zwischen dem “Abrechnungsfehler”, also jenem Eurobetrag, der sich in der Prüfung als fehlerhaft herausstellt, und dem “Zuweisungsfehler”, also jenem Betrag, den die Krankenkasse im Risikopool zu viel erhalten hat, unterschieden werden. Der Zuweisungsfehler lässt sich im Wesentlichen direkt aus dem Abrechnungsfehler herleiten und entspricht der Differenz aus den kassenspezifischen Zuweisungen vor und nach Prüfung. Während bei einer Vollprüfung die Summe der Zuweisungsfehler in der Grundgesamtheit direkt beobachtet werden kann, kann eine Stichprobenprüfung diese Summe nur auf dem Wege der Hochrechnung annäherungsweise bestimmen. Es liegt in der Natur der Sache, dass mit der Stichprobenprüfung eine Ungenauigkeit über die Zuweisungsfehler in der Grundgesamtheit einhergeht.

Im Rahmen des morbiditätsorientierten Risikostrukturausgleiches sind bereits zuverlässige Stichproben- und Hochrechnungsmethoden erarbeitet worden, welche im Rahmen der Datenmeldungen im Bereich der Versichertenzeiten und Morbiditätsdaten verwendet werden. Diese können zwar nicht ohne Weiteres auf die Risikopoolprüfung übertragen werden, eine Anlehnung an diese Prüfverfahren ist aber möglich und könnte die Prüfpflichten, die sich aus §20 RSAV ergeben, vereinheitlichen. Deswegen startet dieses Gutachten mit einer kurzen Zusammenfassung der bereits bestehenden Prüfverfahren. In den darauffolgenden Kapiteln werden dann mögliche Anpassungen diskutiert.

2 Zusammenfassung bestehender RSA-Prüfverfahren

Die bereits bestehenden Prüfverfahren basieren auf einer zweistufigen Stichprobe. In der ersten Stufe wird eine einfache Zufallsstichprobe verwendet, um die kassenindividuelle Fehlerquote zu schätzen. Übersteigt diese einen festgelegten Schwellenwert, wird in der zweiten Stufe eine zusätzliche geschichtete Zufallsstichprobe gezogen. Im Gutachten zu den bestehenden Prüfverfahren wird der kassenindividuelle Stichprobenumfang (n) für die erste Stufe

mittels der folgenden Stichprobenformel bestimmt:

$$n = \frac{\frac{u^2}{\epsilon^2}\frac{1-p}{p}}{1 + \frac{1}{N}\frac{u^2}{\epsilon^2}\frac{1-p}{p}} \tag{1}$$

wobei $u = 1{,}96$ dem 97,5%-Quantil der Standardnormalverteilung entspricht, p die zu erwartende Fehlerquote und ϵ die höchstens zugelassene relative Abweichung vom geschätzten Mittelwert festsetzt. Die Kassengröße wird über den Parameter N, welcher der Anzahl der Versicherten entspricht, in der Formel berücksichtigt. Wenn der kassenindividuelle Stichprobenumfang anhand dieser Formel gewählt wird, dann ist sichergestellt, dass der geschätzte Mittelwert mit 95% Wahrscheinlichkeit höchstens um den Faktor ϵ vom wahren Mittelwert in der Grundgesamtheit aller Versicherten einer Krankenkasse abweicht. Gleichung 1 findet sich auch in diversen Statistiklehrbüchern, beispielsweise in Kauermann und Küchenhoff (2011) sowie Quatember (2019).

Um die Qualität einer gezogenen Stichprobe zu bestimmen, werden in den bisherigen Prüfverfahren die Mittelwerte von 10 Charakteristiken (x) der Versicherten in der Stichprobe mit den entsprechenden Mittelwerten in der Grundgesamtheit der Versicherten in der jeweiligen Krankenkasse verglichen. Hierfür werden anhand der folgenden Formel Konfidenzintervalle um die Stichprobenmittelwerte ($\widehat{x}$) gelegt,

$$KI_{95\%} = \widehat{x} \pm u_B \sqrt{\frac{N-n}{N}}\frac{\sigma}{\sqrt{n}} \tag{2}$$

wobei der verwendete kritische Wert mittels Bonferroni-Korrektur für multiples Testen kontrolliert, $u_B = \Phi^{-1}\left(1 - \frac{0{,}05}{2*10}\right) \approx 2{,}81$. Für binäre Charakteristiken ergibt sich die Standardabweichung σ unmittelbar aus dem zugehörigen Mittelwert $\widehat{x}$, da in diesem Falle einfach gilt $\widehat{\sigma} = \sqrt{\widehat{x}(1-\widehat{x})}$.

Sollten hierbei signifikante Unterschiede zwischen der gezogenen Stichprobe und der Grundgesamtheit festgestellt werden, d. h. mindestens einer der Mittelwerte in der Grundgesamtheit nicht im entsprechenden Konfidenzintervall des Stichproben-Mittelwerts sein, dann wird eine neue Stichprobe gezogen. Dieses Qualitätssicherungsverfahren wird so lange wiederholt, bis sich keine signifikanten Unterschiede mehr zwischen Stichprobe und Grundgesamtheit einer einzelnen Krankenkasse finden lassen.

3 Stichprobenmethodik und Hochrechnung

Im Jahr 2021 umfasst der Risikopool alle Versicherten, deren Ausgaben mindestens 100.000€ betragen (im Folgenden werden diese als Hochrisikoversicherte bezeichnet). Die relevante Grundgesamtheit für die Risikopoolprüfung besteht also nur aus Versicherten mit einem deutlich erhöhten Kostenrisiko. Im Hinblick auf die Stichprobenmethodik und den Stichprobenumfang stellt diese deutlich kleinere Grundgesamtheit einen entscheidenden Unterschied zu den bereits existierenden Prüfverfahren dar. Ziel der neuen Risikopoolprüfung ist eine kassenspezifische Abschätzung der Zuweisungsfehler (Y_i) in der Grundgesamtheit der Hochrisi-

koversicherten einer Krankenkasse.[3] Den bestehenden Prüfverfahren folgend soll die Genauigkeit dieser Abschätzung über alle Krankenkassen hinweg gleich gut ausfallen. Dies dient der Prüfgerechtigkeit.

Zur Unterscheidung der verschiedenen Krankenkassen wird im Folgenden der Subindex j verwendet (für $j = 1, 2, \ldots J$, wobei J die Anzahl der gesetzlichen Krankenkassen angibt). In dem hier betrachteten Abrechnungsjahr (2021) ist $J = 96$. Das monetäre Ziel der Risikoprüfung ist die kassenspezifische Summe der Zuweisungsfehler (μ_j) in der Grundgesamtheit ($i = 1, \ldots, N_j$)

$$\mu_j = \sum_{i=1}^{N_j} Y_i \tag{3}$$

sowie die monetäre Fehlerquote $\left(\bar{Y}_j\right)$, d. h. der Zuweisungsfehler pro Hochrisikoversicherten

$$\bar{Y}_j = \frac{1}{N_j} \sum_{i=1}^{N_j} Y_i \tag{4}$$

wobei N_j der Anzahl der Hochrisikoversicherten in Krankenkasse j entspricht. Abhängig von der Größe der betrachteten Krankenkasse variieren die kassenspezifischen Grundgesamtheiten (N_j) zwischen 1 und 14.767 Hochrisikoversicherten. Da die Prüfkapazität beschränkt ist, scheidet eine Vollprüfung aller Krankenkassen aus. Stattdessen soll eine Kombination aus Voll- und Stichprobenprüfung umgesetzt werden. Für die Stichprobenprüfung wird aus der Grundgesamtheit der Hochrisikoversicherten einer Krankenkasse eine einfache Zufallsstichprobe vom Umfang n_j gezogen und nur diese n_j Fälle werden dann geprüft.[4] Basierend auf dieser Stichprobe lässt sich die kassenspezifische Summe der Zuweisungsfehler wie folgt erwartungstreu hochrechnen[5]

$$\widehat{\mu}_j = N_j \frac{1}{n_j} \sum_{i=1}^{n_j} y_i \, . \tag{5}$$

Die monetäre Fehlerquote lässt sich ebenfalls erwartungstreu schätzen

$$\bar{y}_j = \frac{1}{n_j} \sum_{i=1}^{n_j} y_i \, . \tag{6}$$

Es wird im Folgenden angenommen, dass die Größe der Grundgesamtheit für jede Krankenkasse bekannt ist. Das heißt, bei der Hochrechnung herrscht Unsicherheit nur über die

[3] Für den Risikopool sind die personenbezogenen, ausgleichsfähigen Leistungsausgaben (Satzart 703 bzw. 713) grundlegend. Das Verfahren zur Ermittlung der individuellen Abrechnungsfehler ist nicht Gegenstand dieses Gutachtens und wird dem BAS überlassen. Im Folgenden wird angenommen, dass die fehlerhaften ausgleichsfähigen Ausgaben für jeden Hochrisikoversicherten in einer Stichprobe unproblematisch bestimmt werden können.

[4] Nachdem die Grundgesamtheiten in dem neuen Prüfverfahren deutlich homogener sind als in den bestehenden Prüfungen, sind einfache Zufallsstichproben vermutlich ausreichend. Dieser Punkt kann aber mangels Daten zu den Zuweisungsfehlern noch nicht empirisch untersucht werden. Darüber hinaus würde eine geschichtete Zufallsstichprobe bei kleinen und mittleren Krankenkassen zu sehr kleinen Grundgesamtheiten innerhalb der Schichten führen. Diese Punkte sprechen eher gegen geschichtete Zufallsstichproben.

[5] Zur Notation: kleine Buchstaben bezeichnen Beobachtungen in der Stichprobe, wohingegen große Buchstaben Beobachtungen in der Grundgesamtheit bezeichnen.

monetäre Fehlerquote. Die nachfolgende Diskussion kann sich daher auf deren Schätzung beschränken.

Im Allgemeinen sind die Ergebnisse in einer Stichprobe nicht exakt identisch zu den Werten in der Grundgesamtheit ($\widehat{\mu}_j \neq \mu_j$ und $\bar{y}_j \neq \bar{Y}_j$) und variieren selbst auch je nach gezogener Stichprobe. Die Genauigkeit der Hochrechnung muss also ebenfalls abgeschätzt werden und dies kann mit Hilfe eines Konfidenzintervalls für $\bar{y}_j$ erfolgen. Im Allgemeinen sind Konfidenzintervalle nur asymptotisch gültig, das heißt, es wird unter anderem ein hinreichend großer Stichprobenumfang (n_j) vorausgesetzt. Hartung (2009, S. 272) gibt hier $n_j > 60$ vor. Kauermann und Küchenhoff (2011, S. 28) erwähnen den häufig in der Literatur gewählten Wert $n_j \geq 30$, weisen aber darauf hin, dass dieser Wert vor allem bei kleinem N_j sowie möglichen Ausreißern in der Grundgesamtheit mit Vorsicht zu wählen ist. Die Mehrzahl der Prüffälle wird voraussichtlich keine Abrechnungsfehler aufweisen. Das heißt es ist mit einer Vielzahl von Nullen bei den Zuweisungsfehlern zu rechnen. Dies erschwert eine verlässliche Schätzung der monetären Fehlerquote in kleinen Stichproben. Es ist daher für die Prüfgerechtigkeit wichtig, den Übergang zwischen Voll- und Stichprobenprüfung nicht zu niedrig zu setzen. In Kapitel 4 werden verschiedene Schwellenwerte für die Grundgesamtheit verglichen, ab denen von der Vollprüfung zur Stichprobenprüfung gewechselt wird.

Nachdem die Leistungsausgaben im Risikopool unmittelbar zahlungsbegründend sind, ist vom Gesetzgeber immer eine Inrechnungstellung der Zuweisungsfehler vorgesehen.[6] Während bei der Vollprüfung die Summe der Zuweisungsfehler in der Grundgesamtheit direkt beobachtet werden kann, kann eine Stichprobenprüfung diese Summe nur auf dem Wege der Hochrechnung annäherungsweise bestimmen. Es liegt also in der Natur der Sache, dass nur mit der Stichprobenprüfung eine Ungenauigkeit über die Zuweisungsfehler in der Grundgesamtheit einhergeht. Die nachfolgende Diskussion bezieht sich somit nur auf die Krankenkassen in der Stichprobenprüfung.

Im Interesse der Prüfgerechtigkeit ist es beabsichtigt, die Schätzgenauigkeit für alle Krankenkassen in der Stichprobenprüfung möglichst gleich ausfallen zu lassen. Dies kann erreicht werden, wenn die Breite des Konfidenzintervalls für alle Krankenkassen gleich ausfällt. In den bereits bestehenden Prüfverfahren wird dies dahingehend umgesetzt, dass in einer Stufe-1-Stichprobe erst nur die Breite des Konfidenzintervalls für die *fallbezogene* Fehlerquote über die Krankenkassen hinweg konstant gehalten wird.[7] Deutet die fallbezogene Fehlerquote auf eine Inrechnungstellung hin, dann wird nur für die davon betroffenen Krankenkassen eine Stufe-2-Stichprobe gezogen, die dann auch die *monetäre* Fehlerquote berücksichtigt. Im neuen Prüfverfahren ist eine Inrechnungstellung aber immer vorgesehen, so dass ein zweistufiges Stichprobenverfahren also wieder alle Krankenkassen umfassen müsste und somit diesbezüglich keinen Vorteil bietet.

Das methodische Vorgehen zur Bestimmung des Stichprobenumfangs ist recht ähnlich zu

[6] Mit Inrechnungstellung gemäß §20(5) RSAV ist in diesem Gutachten eine Hochrechnung der Zuweisungsfehler gemeint.

[7] Die fallbezogene Fehlerquote ($\widehat{p}$) wird in den bisherigen Prüfverfahren wie folgt geschätzt

$$\widehat{p} = \frac{n_F}{n} \tag{7}$$

wobei n_F die Anzahl der fehlerhaften Fälle in der Stichprobe angibt.

dem bestehenden Prüfverfahren und wird deswegen an dieser Stelle wieder aufgegriffen. Das asymptotische 95%-Konfidenzintervall für die monetäre Fehlerquote hat die folgende Form[8]

$$KI_{95\%} = \bar{y}_j \pm u \sqrt{\mathrm{Var}(\bar{y}_j)} \tag{8}$$

wobei $u = \Phi^{-1}\left(1-\frac{0{,}05}{2}\right) \approx 1{,}96$ dem 97,5%-Quantil der Standardnormalverteilung entspricht.[9] Die Varianz der geschätzten Fehlerquote wird üblicherweise wie folgt berechnet:

$$\begin{aligned}
\mathrm{Var}(\bar{y}_j) &= \frac{1}{n_j}\frac{1}{n_j-1}\sum_{i=1}^{n_j}\left(y_i-\bar{y}_j\right)^2\left(1-\frac{n_j}{N_j}\right) \\
&= \frac{1}{n_j\frac{N_j}{N_j-n_j}}\frac{1}{n_j-1}\sum_{i=1}^{n_j}\left(y_i-\bar{y}_j\right)^2 \\
&= \frac{1}{\widetilde{n}_j}\frac{1}{n_j-1}\sum_{i=1}^{n_j}\left(y_i-\bar{y}_j\right)^2 .
\end{aligned} \tag{9}$$

Die Genauigkeit der Hochrechnung wird also bestimmt durch den Stichprobenumfang (n_j), die geschätzte Varianz der Zuweisungsfehler ($s_j^2 = \frac{1}{n_j-1}\sum_{i=1}^{n_j}\left(y_i-\bar{y}_j\right)^2$) und einer sogenannten Endlichkeitskorrektur ($1-\frac{n_j}{N_j}$). Der Endlichkeitskorrektur kommt in dem neuen Prüfverfahren eine besondere Bedeutung zu, da diese nur in großen Grundgesamtheiten vernachlässigt werden kann (es gilt nur dann nämlich $1-\frac{n}{N} \approx 1$). In kleineren Grundgesamtheiten führt sie allerdings dazu, dass im Hinblick auf die Genauigkeit der Schätzung zwischen dem *tatsächlichen* Stichprobenumfang (n_j) und einem *effektiven* Stichprobenumfang ($\widetilde{n}_j$) unterschieden werden kann. Der Zusammenhang zwischen den beiden Größen kann wie folgt ausgedrückt werden:

$$\widetilde{n}_j = n_j\frac{N_j}{N_j-n_j} . \tag{10}$$

Bei endlichen Grundgesamtheiten ist der effektive Stichprobenumfang immer größer als der tatsächliche Stichprobenumfang ($\widetilde{n}_j > n_j$), da $N_j > N_j - n_j$. Bei gegebenem Stichprobenumfang ist dieser Effekt umso größer, je kleiner die entsprechende Grundgesamtheit ausfällt. Im Hinblick auf die Schätzgenauigkeit sind in einer kleinen Krankenkasse n Beobachtungen also implizit mehr "wert" als dieselbe Anzahl an Beobachtungen in einer großen Krankenkasse. Beispielsweise liegt für eine Stichprobe mit 10 Beobachtungen in einer Krankenkasse mit $N = 10.000$ Hochrisikoversicherten die effektive Stichprobengröße bei $\widetilde{n} = 10{,}01$. Werden bei einer kleinen Krankenkasse mit nur $N = 50$ Hochrisikoversicherten allerdings 10 Beobachtungen gezogen, so liegt die effektive Stichprobengröße bereits bei $\widetilde{n} = 12{,}50$.

Das asymptotische 95%-Konfidenzintervall für die monetäre Fehlerquote hat letztendlich die

[8] Durch Multiplikation mit N_j kann das Konfidenzintervall für die Hochrechnung ($\widehat{\mu}_j$) aus dem Konfidenzintervall der monetären Fehlerquote berechnet werden.

[9] In der Literatur wird häufig vorgeschlagen, dass Quantil der Standardnormalverteilung durch das entsprechende Quantil der t-Verteilung zu ersetzen. Kauermann und Küchenhoff (2011) folgend, sehen wir hiervon ab, da die t-Verteilung eine Normalverteilung der Zuweisungsfehler voraussetzen würde. Für Stichproben mit mehr als 30 Beobachtungen ist der Unterschied zwischen t-Verteilung und Standardnormalverteilung ohnehin meist vernachlässigbar.

folgende Form

$$KI_{95\%} = \bar{y}_j \left(1 \pm \frac{u}{\bar{y}_j} \frac{s_j}{\sqrt{\widetilde{n}_j}}\right) \tag{11}$$

wobei $\epsilon = \frac{u}{\bar{y}_j} \frac{s_j}{\sqrt{\widetilde{n}_j}}$ ein Maß für die relative Schätzgenauigkeit der monetären Fehlerquote (und somit der späteren Hochrechnung) darstellt. Durch Umformen lässt sich der Zusammenhang zwischen dem vorgegebenen Schätzfehler (ϵ) und dem effektiven Stichprobenumfang wie folgt abbilden[10]

$$\widetilde{n}_j = \frac{u^2}{\epsilon^2} \frac{s_j^2}{\bar{y}_j^2} \tag{12}$$

wobei u wieder dem 97,5%-Quantil der Standardnormalverteilung entspricht. Die Wahrscheinlichkeit, dass die geschätzte monetäre Fehlerquote ($\bar{y}_j$) um mehr als $\epsilon \cdot 100\%$ von der unbekannten monetären Fehlerquote ($\bar{Y}_j$) in der Grundgesamtheit abweicht, soll somit höchstens 5% betragen. Eine zentrale Größe in diesem Zusammenhang ist der Variationskoeffizient $V_j = \frac{s_j}{\bar{y}_j}$, welcher das Verhältnis von Standardabweichung und arithmetischen Mittel misst. Der kassenspezifische Variationskoeffizient muss somit im Vorfeld der Stichprobenziehung bekannt sein. Dieser kann beispielsweise aus dem vorangegangenen Prüfzyklus ermittelt werden.[11] Durch Umformen von Gleichung 10 ergibt sich

$$n_j = \frac{\widetilde{n}_j}{1 + \frac{\widetilde{n}_j}{N_j}} \tag{13}$$

wobei sich letztendlich durch Einsetzen von (12) die gewünschte Schätzgenauigkeit bei der Berechnung des kassenspezifischen Stichprobenumfangs berücksichtigen lässt

$$n_j = \frac{\frac{u^2}{\epsilon^2} \frac{s_j^2}{\bar{y}_j^2}}{1 + \frac{1}{N_j} \frac{u^2}{\epsilon^2} \frac{s_j^2}{\bar{y}_j^2}} \,. \tag{14}$$

Krankenkassen mit gleicher Grundgesamtheit (N_j) und gleichem Variationskoeffizienten (V_j) bekommen also denselben Stichprobenumfang zugewiesen. Andernfalls gilt im Allgemeinen, je größer die Grundgesamtheit und je größer der Variationskoeffizient, desto größer muss der Stichprobenumfang ausfallen, um dieselbe Schätzgenauigkeit zu erhalten. Dies wird erreicht, indem für jede Krankenkasse j eine einfache Zufallsstichprobe im Umfang n_j gezogen wird. Nur diese n_j Fälle werden dann im Rahmen der Stichprobenprüfung kontrolliert. Mit dieser Stichprobe wird die monetäre Fehlerquote geschätzt. Die Hochrechnung erfolgt durch Multiplikation mit der Anzahl der Hochrisikoversicherten N_j.

Um einen Vergleich mit den bestehenden Prüfverfahren zu vereinfachen, wird im Folgenden nochmal kurz die Bestimmung des Stichprobenumfangs in der Stufe-1-Stichprobe wiederholt. In den bestehenden Prüfverfahren wird nur die Schätzgenauigkeit der fallbezogenen

[10] Dies ist ein Standardvorgehen zur Bestimmung des Stichprobenumfangs (siehe beispielsweise S. 275f. in Hartung, 2009).

[11] Für den ersten Prüfzyklus bieten sich möglicherweise Ergebnisse aus einer Pilotprüfung an und/oder ein Konstanthalten der Schätzgenauigkeit für die fallbezogene Fehlerquote (wie in den bereits bestehenden Prüfverfahren).

Fehlerquote konstant gehalten. Der Variationskoeffizient der fallbezogenen Fehlerquote ist in diesem Fall

$$V_j = \sqrt{\frac{1-\widehat{p}_j}{\widehat{p}_j}} \tag{15}$$

woraus sich die Formeln zur Stichprobenbestimmung im bisherigen Prüfverfahren als Spezialfall von Gleichung (12) und (13) ergeben

$$\widetilde{n}_j = \frac{u^2}{\epsilon^2}\frac{1-\widehat{p}_j}{\widehat{p}_j} \tag{16}$$

und sich somit der folgende Stichprobenumfang herleiten lässt

$$n_j = \frac{\frac{u^2}{\epsilon^2}\frac{1-\widehat{p}_j}{\widehat{p}_j}}{1+\frac{1}{N_j}\frac{u^2}{\epsilon^2}\frac{1-\widehat{p}_j}{\widehat{p}_j}}\,. \tag{17}$$

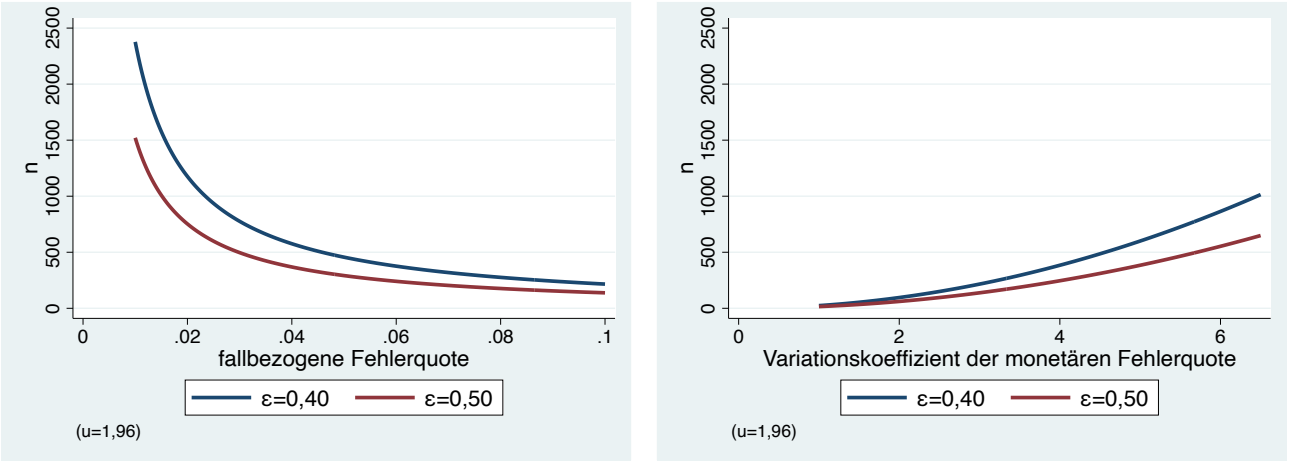
Abbildung 1 stellt den notwendigen Stichprobenumfang als Funktion der fallbezogenen bzw. der monetären Fehlerquote dar. In beiden Darstellungen wird die Endlichkeitskorrektur vernachlässigt (d.h. es wird eine große Krankenkasse mit $N_j \to \infty$ betrachtet, so dass $n_j = \widetilde{n}_j$). Während in Abbildung 1a die Schätzgenauigkeit der fallbezogenen Fehlerquote mit Hilfe von Gleichung (16) konstant gehalten wird, ist dies in 1b für die monetäre Fehlerquote der Fall (basierend auf Gleichung 12). Gerade wenn die fallbezogene Fehlerquote niedrig erwartet wird, ist ein deutlicher Anstieg im notwendigen Stichprobenumfang erforderlich, da der Variationskoeffizient der fallbezogenen Fehlerquote dann sehr hoch ausfällt (siehe Gleichung 15). Der positive Zusammenhang zwischen Variationskoeffizient und Stichprobenumfang ist auch in Abbildung 1b dargestellt. Es wird erwartet, dass der Variationskoeffizient der monetären Fehlerquote je nach Krankenkasse zwischen 2 und 6 liegt (in Kapitel 4 werden die Annahmen hinter diesen Zahlen ausführlich erläutert). Der erste Prüfzyklus bzw. eine Pilotprüfung für unterschiedlich große Krankenkassen kann mehr Gewissheit über die kassenspezifischen Variationskoeffizienten schaffen.

4 Simulation

4.1 Simulationsdesign

In diesem Kapitel soll das Verfahren zur Bestimmung der kassenspezifischen Stichprobengröße mit Hilfe einer Simulation illustriert werden, und zwar unter besonderer Berücksichtigung möglicher Kapazitätsbeschränkungen, der Kombination aus Voll- und Stichprobenprüfung, und der Prüfgerechtigkeit in der Stichprobenprüfung.

Die Simulation erfolgt im Hinblick auf die tatsächliche Krankenkassenstruktur im betrachteten Abrechnungsjahr. Ein Datensatz zur Krankenkassenstruktur für das Jahr 2021 wurde zur Verfügung gestellt, welcher analog zur Verfahrensbestimmung nach §14(4) RSAV erstellt wurde und die Hochrisikofälle, die bei einer einzelnen Krankenkasse Ausgaben über dem



(a) (b)

Abbildung 1: Notwendiger Stichprobenumfang (große Krankenkassen)

Schwellenwert von 100.000€ aufweisen, beinhaltet. Hierin enthalten sind 110.143 Hochrisikoversicherte verteilt auf 96 unterschiedlich große Krankenkassen. Die Datensatzstruktur enthält kleine Krankenkassen mit $N_j = 1$ und große mit $N_j = 14.767$. Die Median-Krankenkasse hat 130 Hochrisikoversicherte, d.h. 50% der Krankenkassen haben bis zu 130 potentielle Prüffälle und die verbleibenden 50% haben mehr als 130 Fälle. Im arithmetischen Mittel sind es 1.147 Hochrisikoversicherte pro Krankenkasse. Abbildung 2 verdeutlicht, dass es viele kleine Krankenkassen gibt, in denen die Anzahl der Hochrisikoversicherten sehr niedrig ausfallen kann. Auch für diese Krankenkassen ist eine kassenspezifische Fehlerquote zu bestimmen. Für einige hiervon wird dies verlässlich nur über eine Vollprüfung möglich sein.

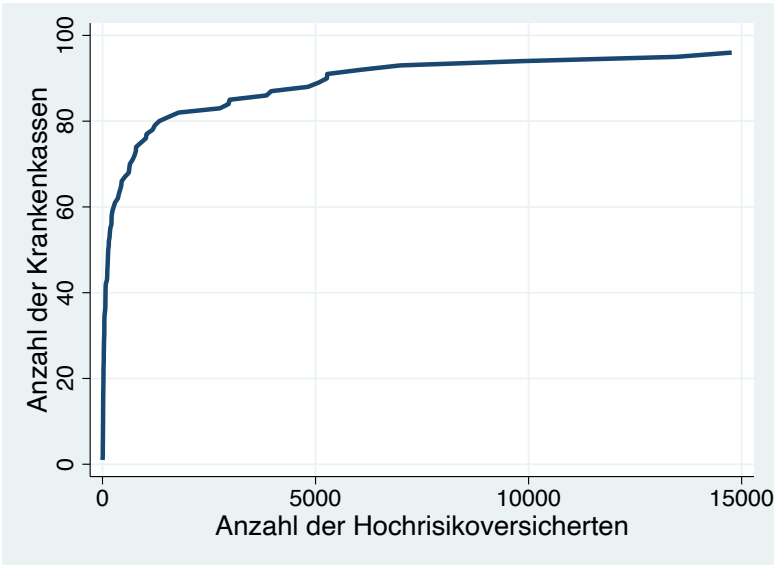


Abbildung 2: Verteilung der Hochrisikoversicherten

Nachdem eine Pilotprüfung noch nicht durchgeführt wurde, liegt zum Zeitpunkt des Gutach-

tens noch kein historischer Datensatz zu den Zuweisungsfehlern im Risikopool vor. Die nachfolgende Diskussion ist also auf Annahmen über die Zuweisungsfehler angewiesen. Die berichteten Ergebnisse sollten somit allenfalls als grobe Illustration der Zusammenhänge verstanden werden. Eine Wiederholung der Simulation mit Ergebnissen aus einer Pilotprüfung ist zur Optimierung des Stichprobenverfahrens angeraten.

Auf Grundlage einer Einschätzung des Arbeitskreises Organisation der Prüfdienste ist davon auszugehen, dass die Abrechnungsfehler proportional zu den tatsächlichen Ausgaben sind. Der zur Verfügung gestellte Datensatz enthält die tatsächlichen Ausgaben für jeden der 110.143 Hochrisikoversicherten. In Absprache mit dem BAS wird der folgende Zusammenhang zwischen monetärem Abrechnungsfehler (F) und tatsächlichen Ausgaben (C) angenommen

$$F_i = 0{,}01\, C_i\, s_i \ . \tag{18}$$

Mit dieser Formel wird dem Umstand Rechnung getragen, dass höhere Ausgaben vermutlich mit einem größeren Abrechnungsfehler einhergehen (circa 1% der Ausgaben werden fehlerhaft erwartet). s ist eine Indikatorvariable mit

$$s_i = \begin{cases} 1 & \text{falls die Abrechnung } i \text{ monetär fehlerhaft ist} \\ 0 & \text{falls die Abrechnung } i \text{ monetär korrekt ist} \,. \end{cases} \tag{19}$$

Der Arbeitskreis Organisation der Prüfdienste erwartet, dass ungefähr 12% der Prüffälle monetär fehlerhaft sind, $Pr(s_i = 1) = 0.12$. Aus diesen Annahmen folgt, dass die Fehlerwahrscheinlichkeit nicht mit den tatsächlichen Ausgaben variiert. In der Tat ist es aber beispielsweise denkbar, dass die fallbezogene Fehlerquote höher ausfällt, wenn die tatsächlichen Ausgaben besonders hoch ausfallen. Ebenfalls einher mit diesen Annahmen geht, dass die Abrechnungsfehler über die Krankenkassen hinweg nur in dem Maße variieren können, indem die tatsächlichen Ausgaben zwischen den Krankenkassen schwanken. Die Proportionalität zwischen Abrechnungsfehler und tatsächlichen Ausgaben ist eine entscheidende Annahme, die mit den zur Verfügung gestellten Daten nicht überprüft werden konnte. Weicht der wahre (aber unbeobachtete) Variationskoeffizient von diesen Annahmen ab, dann hat dies Auswirkungen auf den notwendigen Stichprobenumfang und damit auch auf die Schätzgenauigkeit. Im Rahmen einer Pilotprüfung sollten diese Annahmen empirisch untersucht werden.

Der Zuweisungsfehler (Y_i) ist die Differenz zwischen Zuweisung vor Prüfung und Zuweisung nach Prüfung und lässt sich in der Simulation wie folgt berechnen

$$Y_i = 0{,}8(C_i - 100.000) \ - \ \mathsf{max}\big(0\,;\, 0{,}8(C_i - F_i - 100.000)\big) \tag{20}$$

wobei die Maximum-Funktion im zweiten Term dem Umstand Rechnung trägt, dass die Zuweisungen nie negativ sein sollen.

Für die tatsächliche Krankenkassenstruktur werden unter Verwendung von Gleichung (18) zufällige Abrechnungsfehler in den Grundgesamtheiten (N_j) erzeugt und die damit einhergehenden Zuweisungsfehler berechnet. Die simulierte monetäre Fehlerquote ($\bar{Y} = \frac{1}{N_j} \sum_{i=1}^{N_j} Y_i$) stellt somit das Ziel der Risikoprüfung dar. Ihre Schätzung bildet die Grundlage für eine verlässliche Hochrechnung. In der Simulation sind 13.346 Fälle fehlerhaft (12,12%). Im Mittel sind

dabei 1.405€ zu viel zugewiesen worden (Median: 1.117€). Der größte Zuweisungsfehler liegt bei 21.390€. Der Mittelwert der kassenspezifischen monetären Fehlerquote liegt in der Simulation bei 175,3€ (Median: 166,6€). Bei vier Krankenkassen liegt die simulierte Fehlerquote bei 0€. Insgesamt wären unter diesen Annahmen 18.753.014€ fälschlicherweise zugewiesen worden.

4.2 Schwellenwert zwischen Voll- und Stichprobenprüfung

Als erstes soll untersucht werden, welche Prüfgenauigkeit mit unterschiedlichen Kapazitätsbeschränkung einhergeht. Hierfür wird die gewünschte Kapazität auf die 96 Krankenkassen so verteilt, dass die fallbezogene Fehlerquote für alle Krankenkassen einheitlich gut geschätzt werden kann.[12] Die Beschränkung auf eine gewünschte Prüfkapazität (K) lässt sich in diesem Fall wie folgt modellieren

$$K = \sum_{j=1}^{J} n_j = \sum_{j=1}^{J} \min\left(\frac{\widetilde{n}}{1 + \frac{\widetilde{n}}{N_j}} \; ; \; N_j \right) \tag{21}$$

wobei die zweite Gleichung den Zusammenhang zwischen tatsächlicher und effektiver Stichprobengröße ausnutzt (vgl. Gleichung 13). Wichtig ist anzumerken, dass der effektive Stichprobenumfang $\widetilde{n}$ nun nicht mehr kassenspezifisch ist. Die fallbezogene Fehlerquote kann somit unabhängig von der Krankenkassengröße gleich gut geschätzt werden (unter der Annahme, dass die wahre fallbezogene Fehlerquote über die Krankenkassen tatsächlich konstant ist). Dies entspricht der Bestimmung der Stichprobengröße im bestehenden Prüfverfahren ergänzt um eine mögliche Kapazitätsbeschränkung.

Die Minimum-Funktion in (21) trägt dem Umstand Rechnung, dass in einer Krankenkasse j die Stichprobe (n_j) höchstens so groß wie ihre Grundgesamtheit (N_j) sein kann (dies entspricht dann einer Vollerhebung). Falls also $\frac{\widetilde{n}}{1+\widetilde{n}/N_j} > N_j$, dann wird bei Krankenkasse j eine Vollerhebung durchgeführt. Die verbliebene Prüfkapazität für Krankenkasse j steht dann für andere Prüfungen zur Verfügung. Die erreichbare effektive Stichprobengröße ($\widetilde{n}$) hängt von den Krankenkassengrößen (also der Verteilung von N_j) sowie der gewünschten Prüfkapazität ab und kann durch Lösen von Gleichung (21) bestimmt werden. Nach Bestimmung von $\widetilde{n}$ kann dann n_j für jede Krankenkasse so gewählt werden, dass die gewünschte Kapazitätsbeschränkung nicht überschritten wird. Die Minimum-Funktion modelliert hierbei implizit den Übergang zwischen Vollprüfung und Stichprobenprüfung.

Tabelle 1 gibt den erreichbaren effektiven Stichprobenumfang $\widetilde{n}$ für unterschiedliche Prüfkapazitäten an. Darüber hinaus sind in der zweiten Zeile auch die Stichprobengrößen angegeben, die man erreichen würde, falls man die Prüfkapazität unabhängig von der Krankenkassengröße auf die Krankenkassen verteilen würde ($K/96$). Verbunden mit dem effektiven Stichprobenumfang ist ein höchstens zugelassener Schätzfehler ϵ, welcher mit einer Wahr-

[12] Aus Vereinfachungsgründen wird dies nur für die fallbezogene Fehlerquote bestimmt. Diese ist in der Simulation konstant und somit ist auch der Variationskoeffizient der fallbezogenen Fehlerquote für alle Krankenkassen konstant. Für die monetäre Fehlerquote, die ja das eigentliche Ziel darstellt, ist dies weder in der Simulation der Fall, noch ist zu erwarten, dass dies für die wahre monetäre Fehlerquote zutrifft.

scheinlichkeit von 95% eingehalten wird. Wie zu erwarten, fällt dieser Schätzfehler mit zunehmender Prüfkapazität niedriger aus.[13]

Tabelle 1: Kapazitätsbeschränkungen und implizite Schwellenwerte zur Stichprobenprüfung

K	4.000	8.000	12.000	16.000
$\widetilde{n}$	68,92	176,31	317,14	491,62
$(K/96)$	(41,67)	(83,33)	(125,00)	(166,67)
	Höchstens zugelassener Schätzfehler (ϵ)			
$p = 0.05$	1,029	0,643	0,480	0,385
$p = 0.10$	0,708	0,443	0,330	0,265
$p = 0.15$	0,562	0,351	0,262	0,210
Schwellenwert Vollprüfung	1	9	12	14

Tabelle 1 zeigt auch die implizit ermittelten Schwellenwerte zur Vollprüfung für die betrachteten Szenarien hinsichtlich der gewünschten Kapazität K. Krankenkassen bis zu diesem Schwellenwert würden dann im Rahmen einer Vollprüfung untersucht werden. Eine einheitliche Verteilung der Prüfkapazität auf Voll- und Stichprobenprüfung nach Gleichung (21) führt zu sehr niedrigen Schwellenwerten für die Vollprüfung. Beispielsweise würde eine Vollprüfung bei Krankenkassen mit bis zu neun Hochrisikofällen stattfinden, wenn die Kapazität auf $K = 8.000$ festgelegt wird.[14] Bei dieser Prüfkapazität müssten also bereits Krankenkassen mit nur 10 Hochrisikofällen in die Stichprobenprüfung. Selbst bei einer gewünschten Kapazität von $K = 16.000$ Prüfungen müssten immer noch Krankenkassen mit 15 Hochrisikofällen in die Stichprobenprüfung. Diese Werte sind deutlich unter den Schwellenwerten, welche in der Literatur als notwendig angesehen werden, um die asymptotische Gültigkeit von Konfidenzintervallen zu rechtfertigen (siehe die Diskussion in Kapitel 3). Dies spricht dafür, die Festlegung des Schwellenwerts zwischen Voll- und Stichprobenprüfung unabhängig von den Kapazitätsbeschränkungen zu treffen. Der Literatur folgend sollte dieser Schwellenwert bei mindestens 30, besser sogar bei 50 Fällen liegen. Eine abschließende Beurteilung für einen sinnvollen Schwellenwert ist erst bei Vorliegen von verlässlichen Informationen zu den Zuweisungsfehlern im Risikopool möglich.

Tabelle 2 gibt die Anzahl der Krankenkassen und die Anzahl der Hochrisikoversicherten für unterschiedliche Schwellenwerte zur Stichprobenprüfung an. In der Vollprüfung entspricht die Anzahl der Hochrisikoversicherten der Summe der Prüffälle. Es herrscht somit keine Unsicherheit über die monetäre Fehlerquote, d.h. die Zuweisungsfehler können in der Grundgesamtheit beobachtet werden. Bei einem Schwellenwert von 50 wären beispielsweise 34 Krankenkassen mit zusammen 767 Hochrisikofällen in der Vollprüfung zu prüfen. Aus Abbildung 3 kann die Anzahl der Krankenkassen in der Vollprüfung für unterschiedliche Schwellenwerte abgelesen werden.

[13] Bei diesen Ergebnissen handelt es sich um theoretische Größen, welche bei gegebenem Simulationsdesign (d. h. Krankenkassengrößen, Kapazität und erwarteter Fehlerwahrscheinlichkeit) konstant sind (d.h. unabhängig von einer gezogenen Stichprobe). Der höchstens zugelassene Schätzfehler ϵ hängt hierbei auch von der zu erwartenden Fehlerwahrscheinlichkeit p ab (vgl. Gleichung 16)

[14] Dieser Wert wirkt zwar verhältnismäßig niedrig, es ist aber in diesem Szenario zu bedenken, dass bei 96 Krankenkassen ein "Sockel" von neun Prüfungen in jeder Krankenkasse bereits 10,80% der Prüfkapazität verbrauchen würde.

Tabelle 2: Explizite Festlegung des Schwellenwertes zur Stichprobenprüfung

Schwellenwert	$N_j < 30$	$N_j \geq 30$	$N_j < 50$	$N_j \geq 50$	$N_j < 100$	$N_j \geq 100$
Anzahl der Krankenkassen	23	73	34	62	42	54
Hochrisikoversicherten:						
Gesamtzahl der Fälle	349	109.794	767	109.376	1.312	108.831
Pro Krankenkasse (Median)	16	262	19	445	24	636

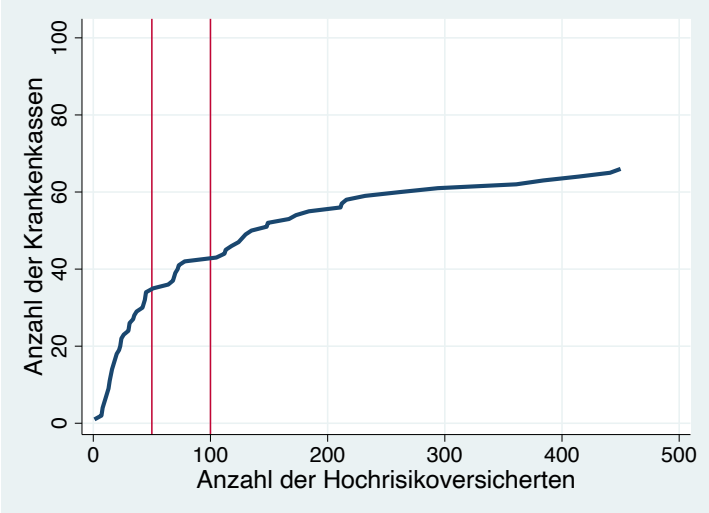


Abbildung 3: Anzahl der Krankenkassen in der Vollprüfung

4.3 Simulationsergebnisse in der Stichprobenprüfung

In diesem Abschnitt soll die Schätzgenauigkeit empirisch mit Hilfe von Simulationen untersucht werden. Hierfür wird der Datensatz auf die Krankenkassen in der Stichprobenprüfung eingeschränkt. Die Ergebnisse werden für $N_j \geq 50$ als Schwellenwert berichtet. Die Schätzqualität in der Stichprobenprüfung wird den bestehenden Prüfverfahren folgend auf $\epsilon = 0{,}40$ festgelegt.[15] Damit einher geht eine notwendige Prüfkapazität in der Stichprobenprüfung. Die Gesamtkapazität der Prüfungen ergibt sich dann als Summe der Stichproben- und Vollprüfungen. Für die Bestimmung der Stichprobengröße sind Informationen über die kassenspezifischen Variationskoeffizienten notwendig. In der Anwendung sind diese natürlich zuerst unbekannt. Im Folgenden werden zwei mögliche Ansätze diskutiert, um diese im Vorfeld der Stichprobenbestimmung festzulegen. Im ersten Ansatz werden die kassenspezifischen Variationskoeffizienten direkt für die Zuweisungsfehler ermittelt (beispielsweise aus dem vorangegangen Prüfzyklus).[16] Dieser Ansatz erlaubt es, den Stichprobenumfang so festzulegen, dass die monetäre Fehlerquote für alle Krankenkassen einheitlich gut geschätzt werden kann. Im zweiten Ansatz wird nur Vorwissen über die GKV-weite fallbezogene Fehlerquote verwendet. Hierdurch nimmt man implizit an, dass die Variationskoeffizienten in allen Krankenkassen gleich ausfallen. Der Stichprobenumfang wird dann so festgelegt, dass die fall-

[15] Weitere Simulationsergebnisse für unterschiedliche Schwellenwerte und Schätzgenauigkeiten finden sich im Anhang.

[16] Dies wird in Simulation dadurch abgebildet, dass die Information aus der vorangegangenen Stichprobe verwendet werden.

bezogene Fehlerquote einheitlich gut geschätzt werden kann. Da das Ziel aber trotzdem die monetäre Fehlerquote ist, wird der zweite Ansatz die Genauigkeit dieser Schätzung natürlich nicht konstant halten können. Er könnte aber für die Pilotprüfung und/oder den ersten Prüfzyklus hilfreich sein und entspricht im Wesentlichen dem bereits bestehenden Prüfverfahren in der Stufe-1-Stichprobe.

Um die Qualität des Verfahrens zu evaluieren, wird die Stichprobenziehung insgesamt 1000-mal wiederholt und für jede dieser Stichproben die monetäre Fehlerquote sowie deren (asymptotischer) Standardfehler als Maß für die Schätzgenauigkeit berechnet. Nach Gleichung (11) kann daraus dann das beobachtete $\widehat{\epsilon}_r$ für jede der 1000 Wiederholungen berechnet werden. Daraus lässt sich dann die mittlere Schätzgenauigkeit $(\widehat{\epsilon} = \frac{1}{1000}\sum_{r=1}^{1000}\widehat{\epsilon}_r)$ für alle Krankenkassen in der Stichprobenprüfung bestimmen.

Abbildung 4 zeigt den ermittelten Stichprobenumfang unter Verwendung des ersten Ansatzes – abhängig von der Krankenkassengröße. Die Schwankungen bei im Wesentlichen gleicher Krankenkassengröße resultieren aus den unterschiedlichen Variationskoeffizienten. Um die asymptotische Gültigkeit der Konfidenzintervalle zu rechtfertigen, sollte der Stichprobenumfang auch in der Stichprobenprüfung den Schwellenwert von 50 Prüffällen pro Krankenversicherung nicht unterschreiten. Der Stichprobenumfang kann andernfalls auf diesen Mindestwert angepasst werden. Abbildung 5 illustriert, wiederum abhängig von der Krankenkassengröße, die mittlere Schätzgenauigkeit $(\widehat{\epsilon})$ für alle Krankenkassen in der Stichprobenprüfung. Bei dem ersten Ansatz kann die Schätzgenauigkeit hinreichend gut kontrolliert werden (siehe Abbildung 5). Abbildung 6 und 7 zeigen die Ergebnisse für den zweiten Ansatz, in welchem die GKV-weite Fehlerquote zur Bestimmung des Variationskoeffizienten verwendet wird (beispielsweise ebenfalls aus einem vorangegangen Prüfzyklus ermittelt). Im Gegensatz zum ersten Ansatz wird hierin die Schätzgenauigkeit der fallbezogenen Fehlerquote konstant gehalten, nicht jedoch die Schätzgenauigkeit der monetären Fehlerquote. Dies ist deutlich aus Abbildung 7 ersichtlich. Tabelle 3 gibt die Gesamtzahl der Prüffälle aus Voll- und Stichprobenprüfung an. Die Gesamtzahl der Prüffälle in der Stichprobenprüfung kann geringfügig höher ausfallen wenn auch hier in jeder Krankenkasse mindestens 50 Prüffälle geprüft werden sollen.

Tabelle 3: Summe der Prüffälle in der Voll- und Stichprobenprüfung

	Vollprüfung	Stichprobenprüfung			Summe		
		Minimum	Mittelwert	Maximum	Minimum	Mittelwert	Maximum
1. Ansatz $N_j \geq 50$	767	8.288	9.286	10.790	9.055	10.053	11.557
2. Ansatz $N_j \geq 50$	767	6.677	7.201	7.762	7.444	7.968	8.529

Sowohl die theoretischen Überlegungen aus Kapitel 3 als auch die Simulationsergebnisse bestätigen also, dass die stichprobenbasierten Prüfverfahren Krankenkassen mit unterschiedlicher Größe im Hinblick auf die Qualität der Schätzung möglichst gleich behandeln. Von einer möglichen Erhöhung der Prüfkapazität würden somit alle Krankenkassen unabhängig von deren Größe gleichermaßen profitieren.

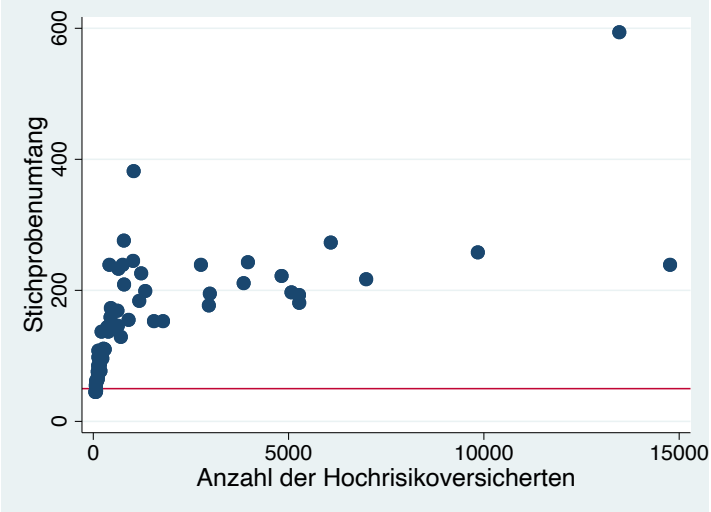


Abbildung 4: Ermittelter Stichprobenumfang in einer Wiederholung (1. Ansatz)

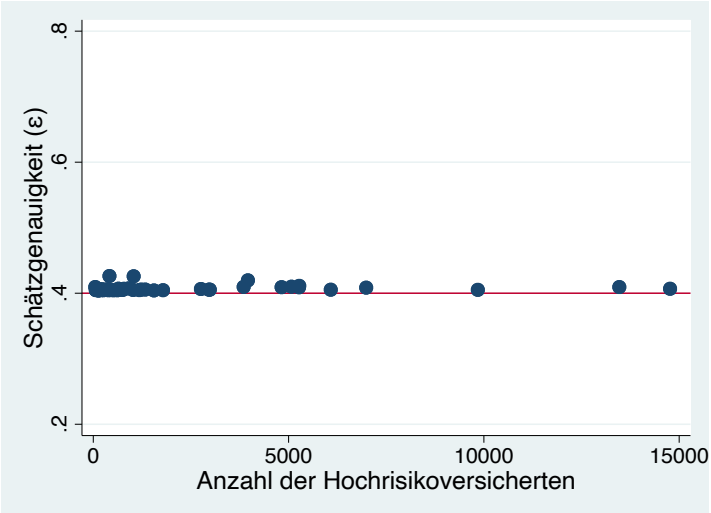


Abbildung 5: Schätzgenauigkeit der monetären Fehlerquote in der Simulation (1. Ansatz)

5 Stichprobengüte

Aus Gründen der Prüfgerechtigkeit ist es beabsichtigt Stichproben auszuschließen, die in auffälliger Weise nicht "repräsentativ" für die Grundgesamtheit einer Krankenkasse erscheinen. Es sei darauf hingewiesen, dass solch ein Verfahren Auswirkungen auf das asymptotische Verhalten der nachfolgenden Schätzung haben kann. Im Interesse der Prüfgerechtigkeit kann es aber trotzdem angebracht sein. In den Simulationen in Kapitel 4 wurde auf die Bestimmung der Stichprobengüte jedoch verzichtet.

In den bereits bestehenden Prüfverfahren wird die Stichprobengüte anhand von 10 Versichertencharakteristiken mit Hilfe von Mittelwertvergleichen überprüft (siehe auch Kapitel 2). Hierfür werden bisher nur um die Mittelwerte aus der Stichprobe Konfidenzintervalle gelegt, nicht aber um die entsprechenden Mittelwerte aus der Grundgesamtheit. Dies kann in den

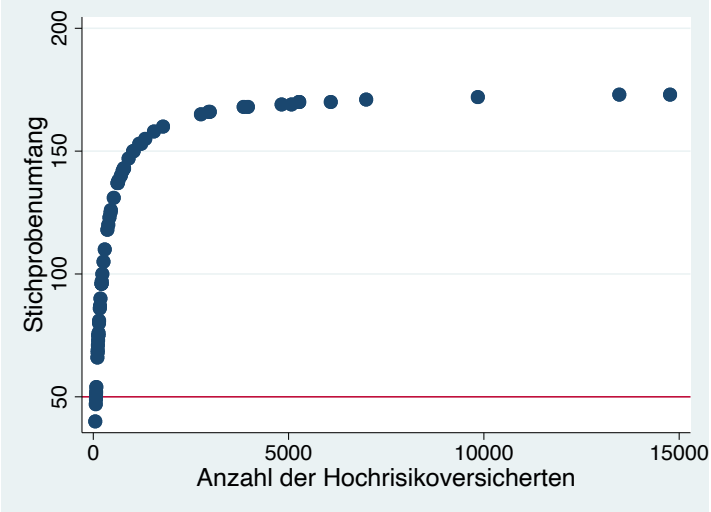


Abbildung 6: Ermittelter Stichprobenumfang in einer Wiederholung (2. Ansatz)

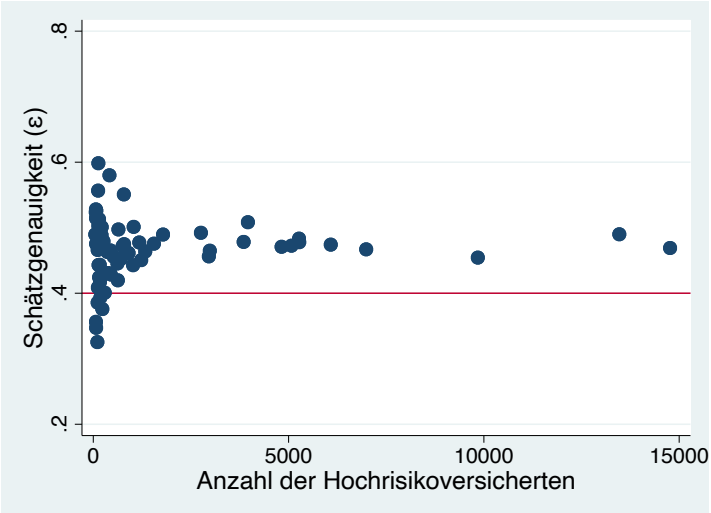


Abbildung 7: Schätzgenauigkeit der monetären Fehlerquote in der Simulation (2. Ansatz)

bestehenden Prüfverfahren damit gerechtfertigt werden, dass die Grundgesamtheit in allen Krankenkassen hinreichend groß ist und somit die Mittelwerte aus der Grundgesamtheit sehr präzise vorhanden sind. Sie verhalten sich also annähernd wie Erwartungswerte. Diese Rechtfertigung fällt aber im neuen Prüfverfahren weg, weil die Grundgesamtheiten selbst sehr klein ausfallen können (so liegt bei den Krankenkassen in der Stichprobenprüfung bei einem Schwellenwert von 50 der Median der Grundgesamtheiten bei 445 Hochrisikofällen pro Krankenkasse).

Die Kontrolle der Stichprobengüte sollte für das neue Prüfverfahren geringfügig verändert werden. Hierfür gibt es mehrere Möglichkeiten – zwei davon werden im Folgenden kurz skizziert. Die erste und dem bisherigen Prüfverfahren ähnlichste Methode ist ein Mittelwertsvergleich zwischen zwei Stichproben bzw. zwei Gruppen, so dass die Unsicherheit der Mittelwerte aus der Stichprobe und aus der verbleibenden Grundgesamtheit berücksichtigt wird.

Dies lässt sich beispielsweise über einfache Regressionen erreichen. Die abhängigen Variablen hierbei sind die 10 Versichertencharakteristiken und die einzige erklärende Variable ist eine Indikatorvariable, welche auf die Beobachtungen in der gezogenen Stichprobe verweist. Die zur Indikatorvariable gehörenden Regressionskoeffizienten entsprechen der Mittelwertdifferenz zwischen den zwei Gruppen (also zwischen den Versicherten, die in der Stichprobe sind, und den Versicherten, die nicht in der Stichprobe sind). Sollten sich diese Differenzen als signifikant herausstellen, dann spricht dies für eine (zufälligerweise) nicht sonderlich repräsentative Stichprobe. In diesem Falle kann, dem bisherigen Prüfverfahren folgend, die Stichprobe verworfen und eine neue Stichprobe gezogen werden. Folgende Punkte sind bei diesem Verfahren zur Stichprobengüte noch zu beachten. Erstens, um das Problem des multiplen Testens zu lösen, ist der kritische Wert für die Signifikanz anzupassen. Bei 10 Charakteristiken kann hierfür $u_B = 2{,}81$ als kritischer Wert gewählt werden (Bonferroni-Korrektur). Zweitens, bei den Regressionen sollten robuste Standardfehler verwendet werden, da einige der Versichertencharakteristiken binär sind.

Die zweite Methode greift ebenfalls auf eine lineare Regression zurück, wobei nun die Indikatorvariable, welche auf die Beobachtungen in der Stichprobe verweist, die abhängige Variable ist. Als erklärende Variablen werden die 10 Versichertencharakteristiken verwendet. Im Anschluss wird mit Hilfe eines F-Tests nach signifikanten Koeffizienten gesucht. Sollten sich diese nicht nachweisen lassen, kann die gezogene Stichprobe als hinreichend gut erachtet werden. Nachdem im durchgeführten F-Test eine gemeinsame Hypothese getestet wird, bedarf es keiner Bonferroni-Korrektur. Dieses zweite Testverfahren hat also den Vorteil, dass das gesetzte Signifikanzniveau nicht deutlich unterschritten wird, während die bis jetzt verwendete Bonferroni-Korrektur als eher konservativ gilt, d. h. die Nullhypothese einer repräsentativen Stichprobe eher selten abgelehnt wird. Dieses Verfahren setzt allerdings eine hinreichend große Grundgesamtheit voraus und würde somit nur für die größeren Krankenkassen in Betracht kommen.

6 Schlussbemerkung

Die entscheidende Unbekannte bei der Bestimmung des kassenspezifischen Stichprobenumfangs ist der kassenspezifische Variationskoeffizient der Zuweisungsfehler. Eine Pilotprüfung und/oder der erste Prüfzyklus ermöglichen es, die (fallbezogene und) die monetäre Fehlerquote sowie die zugehörige Varianz zu ermitteln. Im besten Fall liegen diese Informationen dann für jede Krankenkasse in der Stichprobenprüfung getrennt vor und können zur Berechnung des kassenspezifischen Variationskoeffizienten verwendet werden. Dieser kann dann wiederum benutzt werden, um den passenden Stichprobenumfang für den nächsten Prüfzyklus zu bestimmen.

Nachdem die Grundgesamtheiten in der Risikopoolprüfung deutlich homogener ausfallen als in den bestehenden Prüfverfahren werden auch die möglichen Schichtungsgewinne nicht so deutlich ins Gewicht fallen. Dies ist aber eine empirische Frage und kann mangels Daten vorerst nicht beantwortet werden. Die gewonnenen Daten aus der Pilotprüfung bzw. den ersten Prüfzyklen könnten auch auf potentielle Schichtungsgewinne untersucht werden. Angelehnt an die bestehenden Prüfverfahren sind die versichertenbezogenen Zuweisungen eine mög-

licherweise geeignete Schichtungsvariable.

Literaturverzeichnis

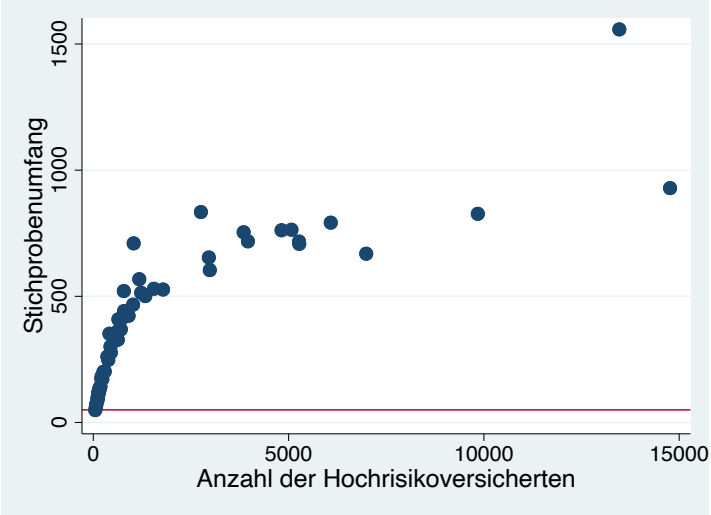
Hartung, Joachim (2009): Statistik. Lehr- und Handbuch der angewandten Statistik, Oldenbourg.

Kauermann, Göran und Küchenhoff, Helmut (2011): Stichproben. Methoden und praktische Umsetzung mit R, Springer.

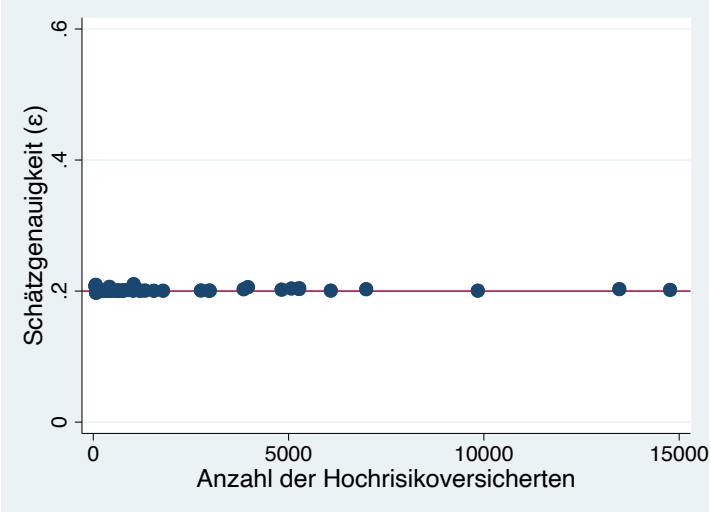
Quatember, Andreas (2019): Datenqualität in Stichprobenerhebungen. Eine verständnisorientierte Einführung in die Survey-Statistik, Springer.

Anhang

Simulationsergebnisse für $N_j \geq 50$ und $\epsilon = 0{,}20$



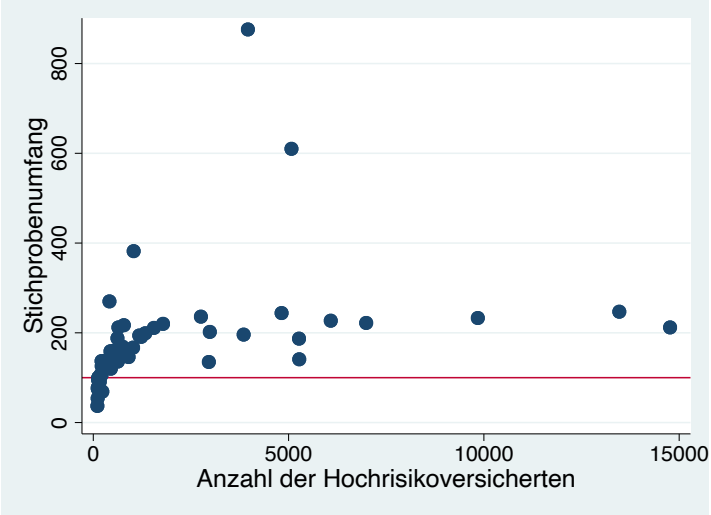
Ermittelter Stichprobenumfang in einer Wiederholung (1. Ansatz)



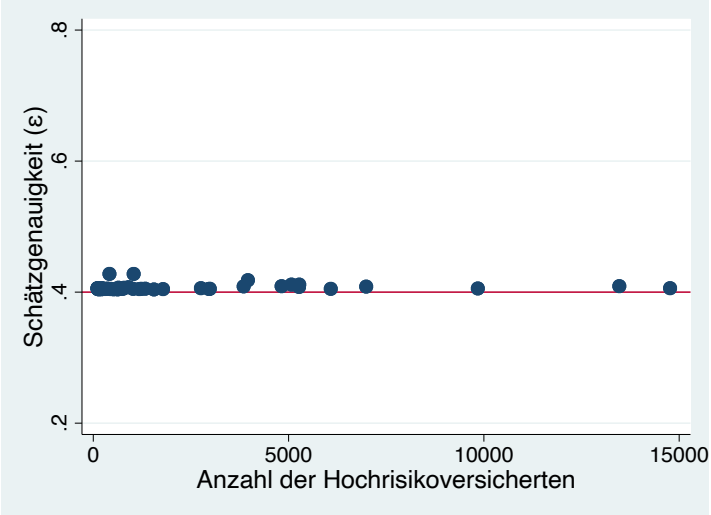
Schätzgenauigkeit der monetären Fehlerquote in der Simulation (1. Ansatz)

Hierfür wären zwischen 22.105 und 26.652 Prüffälle (Mittelwert 23.997) in der Stichprobenprüfung sowie weiterhin 767 Prüffälle in der Vollprüfung zu bearbeiten.

Simulationsergebnisse für $N_j \geq 100$ und $\epsilon = 0{,}40$



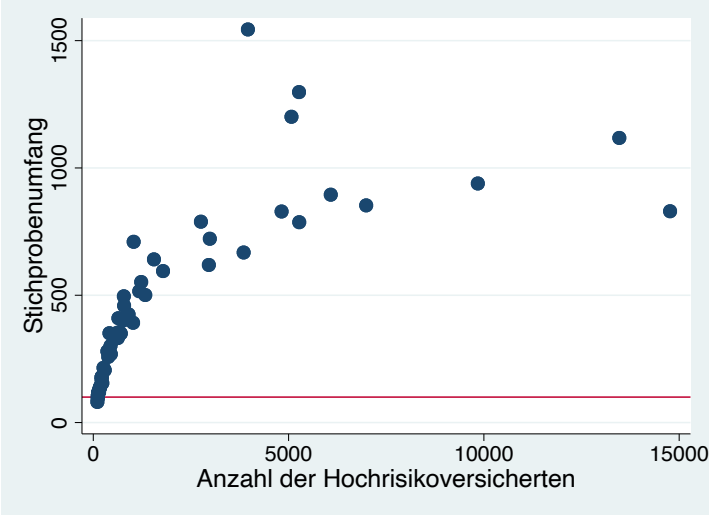
Ermittelter Stichprobenumfang in einer Wiederholung (1. Ansatz)



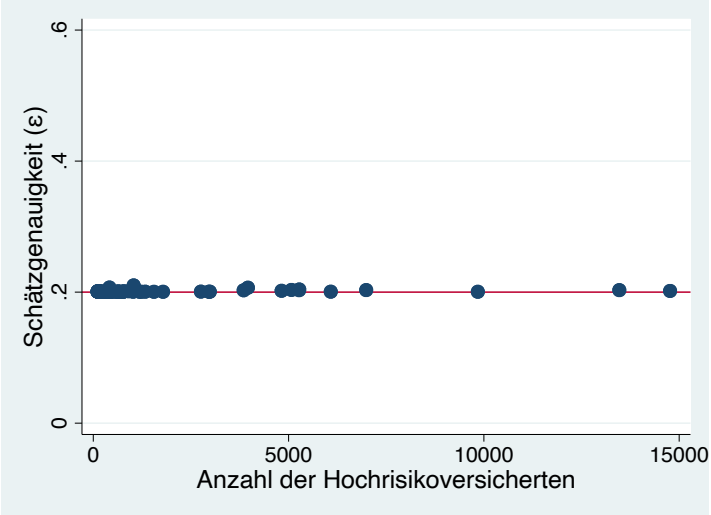
Schätzgenauigkeit der monetären Fehlerquote in der Simulation (1. Ansatz)

Hierfür wären zwischen 7.959 und 10.378 Prüffälle (Mittelwert 8.862) in der Stichprobenprüfung sowie 1.312 Prüffälle in der Vollprüfung zu bearbeiten.

Simulationsergebnisse für $N_j \geq 100$ und $\epsilon = 0{,}20$



Ermittelter Stichprobenumfang in einer Wiederholung (1. Ansatz)



Schätzgenauigkeit der monetären Fehlerquote in der Simulation (1. Ansatz)

Hierfür wären zwischen 21.708 und 26.190 Prüffälle (Mittelwert 23.474) in der Stichprobenprüfung sowie 1.312 Prüffälle in der Vollprüfung zu bearbeiten.