Zi-Paper 22 2022

Methoden zur Gewichtung und deren Anwendung im Zahnärzte-Praxis-Panel

Prof. Dr. Marco Caliendo[1], Chris Kerber[2], Tim Müller-Walden[3], Cosima Obst[4], Markus Leibner[5]

Herausgeber

Zentralinstitut für die
kassenärztliche Versorgung
in der Bundesrepublik Deutschland
Salzufer 8
10587 Berlin
www.zi.de

Autoren

[1] Prof. Dr. Marco Caliendo, Professor für empirische Wirtschaftsforschung, Universität Potsdam
[2] Chris Kerber, wissenschaftlicher Mitarbeiter im Fachbereich Ökonomie
[3] Tim Müller-Walden, wissenschaftlicher Mitarbeiter im Fachbereich Ökonomie
[4] Cosima Obst, wissenschaftliche Mitarbeiterin, Universität Potsdam
[5] Markus Leibner, Fachbereichsleiter Ökonomie

Berlin, den 04.07.2022

ISSN 2199-1480 (online)

Inhaltsverzeichnis

Tabellenverzeichnis

Abbildungsverzeichnis

Abkürzungsverzeichnis

ALLBUS	Allgemeine Bevölkerungsumfrage der Sozialwissenschaften
BAG	Berufsausübungsgemeinschaft
EP	Einzelpraxis
HRF	Hochrechnungsfaktor
IAB	Institut für Arbeitsmarkt- und Berufsforschung
KZBV	Kassenzahnärztliche Bundesvereinigung
KZV	Kassenzahnärztliche Vereinigung
MAR	Missing at Random
MSE	Mean Squared Error
INKAR	Indikatoren und Karten zur Raum- und Stadtentwicklung
SOEP	Sozio-oekonomisches Panel
ZäPP	Zahnärzte-Praxis-Panel
Zi	Zentralinstitut für die kassenärztliche Versorgung in Deutschland
ZiPP	Zi-Praxis-Panel

1 Einleitung

Das *Zahnärzte-Praxis-Panel* (*ZäPP*) ist die Kostenstrukturerhebung der Kassenzahnärztlichen Bundesvereinigung (KZBV). Es handelt sich um eine bundesweite Wiederholungsbefragung von Praxen der vertragszahnärztlichen Versorgung zu Praxisstruktur, erbrachten zahnärztlichen Leistungen sowie Praxisaufwendungen und -einnahmen. Ziel der Erhebung ist die möglichst präzise Abbildung der wirtschaftlichen Situation der zahnärztlichen Praxen und der vertragszahnärztlichen Versorgungstätigkeit in Deutschland.

Im ZäPP werden sämtliche Vertragszahnarztpraxen, die im Befragungszeitraum in Betrieb waren, um Auskunft gebeten. Dabei handelt es sich um rund 38.000 Praxen von denen in den vergangenen Erhebungswellen zwischen 3.300 und 4.500 Praxen teilgenommen haben. Die Teilnahme am ZäPP ist freiwillig, eine Teilnahme- und Auskunftspflicht besteht nicht. Dementsprechend ist eine Selbstselektion der teilnehmenden Praxen im Hinblick auf wichtige Strukturmerkmale möglich. Wenn die Antwortausfälle nicht zufällig sind, sondern systematisch mit den interessierenden Variablen zusammenhängen, liefert die Stichprobe verzerrte Ergebnisse über die Zielpopulation. Um solchen Verzerrungen entgegenzuwirken und die Grundgesamtheit so gut wie möglich zu repräsentieren, werden die Ergebnisse im ZäPP gewichtet.

Die Gewichtung im ZäPP weist einige Besonderheiten auf. Die Daten werden sowohl auf Bundesebene als auch auf Ebene der einzelnen Kassenzahnärztlichen Vereinigungen (KZV) ausgewertet. Auf Bundesebene erfolgt die Gewichtung auf Grundlage von drei Merkmalen, die für die Grundgesamtheit klassiert und kombiniert vorliegen. Bei diesen Merkmalen handelt es sich um die Einnahmen aus vertragszahnärztlicher Tätigkeit (Honorarklassen), die Zugehörigkeit zur jeweils zuständigen Kassenzahnärztlichen Vereinigung (KZV-Zugehörigkeit) und die Organisationsform der Praxen (Einzelpraxis [EP] oder Berufsausübungsgemeinschaft [BAG]). Auf Ebene der Kassenzahnärztlichen Vereinigungen werden der Gewichtung nur die beiden Merkmale Honorarklasse und Organisationsform zugrunde gelegt. In diesem Papier wird das Gewichtungsverfahren im ZäPP vorgestellt. Angesichts rückläufiger Teilnahmezahlen wird außerdem im Rahmen der regelmäßigen Qualitätskontrolle überprüft, ob das angewandte Verfahren weiterhin für das ZäPP geeignet ist oder ob eine Anpassung präzisere Schätzungen ermöglichen würde. Dazu werden zunächst Ursachen für die Verzerrung von Stichproben erläutert und gängige Gewichtungsmethoden beschrieben. Behandelt werden die beiden Poststratifizierungsmethoden *cell weighting* und *raking* sowie eine logistische Regression auf Basis einer Ausfallanalyse. Anschließend werden das ZäPP und ergänzende externe Daten, die für die Ausfallanalyse verwendet werden, vorgestellt. Dabei handelt es sich um Daten aus den Zahnarztregistern und die Umsatzklassenstatistiken der Kassenzahnärztlichen Vereinigung (KZV) sowie ergänzende sozioökonomische Informationen.

Auf Grundlage dieser Daten wird überprüft ob systematische Antwortausfälle vorliegen, die zu einem *nonresponse bias* führen könnten. Bei dieser Ausfallanalyse werden Informationen, die sowohl für Teilnehmer:innen als auch für Nichtteilnehmer:innen vorliegen, dazu verwendet, um die jeweiligen Teilnahmewahrscheinlichkeiten zu bestimmen. Eine Korrelation der Teilnahmewahrscheinlichkeit mit beobachteten Merkmalen deutet auf eine höhere oder niedrigere Teilnahmebereitschaft für Personen mit bestimmten Ausprägungen dieser Merkmale hin. Das hätte zur Folge, dass Subgruppen dieser Merkmale überproportional häufig oder selten in der Stichprobe vertreten sind und damit ein *nonresponse bias* vorliegt.

Anschließend wird das bisherige Gewichtungsverfahren im ZäPP detailliert erläutert. Bislang findet ein *cell weighting*-Ansatz Verwendung, bei dem die Gewichtungsklassen aus der Grundgesamtheit heraus gleichmäßig besetzt werden und jeweils mindestens sechs Beobachtungen enthalten müssen. Da diese Kriterien beeinflussen, wie viele Gewichtungsklassen gebildet werden können und die Anzahl der Gewichtungsklassen die Genauigkeit einer Schätzung beeinflusst, hat ihre Festlegung Einfluss auf die Genauigkeit der Ergebnisse. Mit einer Simulation wird daher unter Berücksichtigung der Besonderheiten des ZäPP bestimmt, wie sich die Mindestbesetzung auf die Schätzung auswirkt und ob die Gewichtungsklassen aus der Grundgesamtheit oder aus der Stichprobe heraus gebildet werden sollen, um die präzisesten Ergebnisse zu liefern.

Die ermittelten Parameter werden dem anschließenden Methodenvergleich zugrunde gelegt, bei dem anhand der Daten aus der Erhebung des ZäPP 2020 ermittelt wird, welches Gewichtungsverfahren bzw. welche Verfahrenskombination die präziseste Schätzung erreicht. Verglichen werden die vorgestellten Methoden *cell*

weighting und *raking* sowie eine logistische Regression auf Grundlage der durchgeführten Ausfallanalyse. Dieser Methodenvergleich ist Grundlage des abschließenden Fazits dieses Papiers.

2. Theorie

Dieses Kapitel liefert einen Einblick in die theoretischen Hintergründe des Verzerrungsproblems. Dabei wird zuerst erläutert, welche Ursachen dazu beitragen können, dass eine Stichprobe nicht repräsentativ ist. Entsprechend den vorliegenden Problemen können dann Lösungsansätze gewählt werden. Im Anschluss werden drei Gewichtungsverfahren als potentielle Lösungsansätze erläutert.

2.1 Verzerrungen der Stichprobe

Bei der Auswertung von Daten ist oftmals das Ziel, eine repräsentative Aussage über die Grundgesamtheit zu treffen. Hierzu ist es notwendig, dass die Grundgesamtheit repräsentativ durch die Stichprobe abgebildet wird. Das kann mithilfe reiner Zufallsstichproben erzielt werden, bei der jede Einheit der Grundgesamtheit dieselbe Wahrscheinlichkeit hat, zufällig in die Stichprobe aufgenommen zu werden. Hierbei gelangen Teilgruppen seltener in die Stichprobe, wenn sie in der Grundgesamtheit bereits selten vertreten sind (z. B. migrierte Personen). Umgekehrt sind große Teilgruppen der Grundgesamtheit auch in der Stichprobe häufiger vertreten. Perfekte Repräsentativität liegt vor, wenn alle Teilgruppen jeweils denselben Anteil in der Grundgesamtheit und in der Stichprobe ausmachen.

Verschiedene Aspekte können bei Erhebungsdaten dazu führen, dass die Repräsentativität der Daten eingeschränkt ist. Ein Verständnis der potentiellen Ursachen kann bei der Wahl des Lösungsansatzes hilfreich sein. Folglich werden in diesem Abschnitt die möglichen Ursachen von Verzerrungen ergründet.

Stichprobendesign: Das Studiendesign spielt eine zentrale Rolle bei der Zusammensetzung der gezogenen Stichprobe. Der Nachteil der reinen Zufallsstichprobe ist, dass kleinere Teilgruppen der Grundgesamtheit entsprechend selten in der Stichprobe vorhanden sind. Durch die geringe Anzahl an Beobachtungseinheiten dieser Teilgruppe ist es wiederum schwierig, belastbare Aussagen über diese Teilgruppe zu treffen. Als Konsequenz erhalten in manchen Studien unterschiedliche Gruppen (Schichten) der Grundgesamtheit unterschiedliche Wahrscheinlichkeiten, Teil der Stichprobe zu werden. Dadurch verliert die Stichprobe allerdings ihre Repräsentativität, weil einzelne Gruppen disproportional oft oder selten vorkommen.

Ein Beispiel hierfür ist das deutsche Sozio-oekonomische Panel (SOEP), eine der größten und am längsten laufenden Haushalts-Panelbefragungen weltweit. Hier sind beispielsweise Migrant:innen überrepräsentiert; ihr Anteil unter den Haushalten des Sozio-oekonomisches Panel (SOEP) ist größer als ihr Anteil unter allen Haushalten der Bundesrepublik Deutschland (Goebel et al., 2019). Als weiteres Beispiel kann das IAB-Betriebspanel aufgeführt werden, welches jährlich durch das Institut für Arbeitsmarkt- und Berufsforschung (IAB) durchgeführt wird. Hier sind überproportional oft kleine Bundesländer und kleine Branchen vertreten (Bachmann, Tschersich, Ellguth, Kohaut & Baier, 2020).

Tabelle 1 Ausschöpfungsquoten im ZäPP

Erhebungswelle	Ausschöpfungsquote
ZäPP 2018	12,7%
ZäPP 2019	9,3%
ZäPP 2020	9,2%

Quellen: Zahnärzte-Praxis-Panel 2018-2020.
Anmerkungen: Die Tabelle zeigt die Ausschöpfungsquoten in % der Jahre 2018-2020.

Nonresponse: Sowohl bei reinen Zufalls- als auch bei geschichteten Stichproben kann es zu Antwortausfällen auf Beobachtungsebene kommen, der *unit nonresponse*. Hierbei ist es Beobachtungseinheiten entweder nicht möglich zu antworten, oder sie entscheiden sich gegen eine Auskunft. Dadurch ist die befragte Stichprobe kleiner als die kontaktierte Stichprobe. Die resultierenden Ausschöpfungsquoten haben in den vergangenen Jahren bei verschiedenen Erhebungsdaten einen Rückgang verzeichnet. So ist z. B. bei der Allgemeinen Bevölkerungsumfrage der Sozialwissenschaften (ALLBUS, Blumenstiel & Gummer, 2015) die Ausschöpfungsquote innerhalb von 6 Jahren von 45% auf 35% gesunken. Auch das ZäPP ist davon nicht verschont geblieben wie Tabelle 1 zeigt. Dabei ist allerdings zu beachten, dass geringe Ausschöpfungsquoten nicht automatisch Verzerrungen verursachen (Groves, 2006; Groves & Peytcheva, 2008), da auch klei-

ne Stichproben repräsentativ sein können. Zentral ist daher die Frage, ob die *unit nonresponse* systematisch auftritt. Falls bestimmte Teilgruppen der Grundgesamtheit systematisch die Antwort verweigern, kann das wiederum zu Verzerrungen führen, dem *nonresponse bias*.

Entsprechend ist für jede Stichprobe eine Einschätzung notwendig, ob ein *nonresponse bias* vorliegt. Hierzu wird nach Möglichkeit die Teilnahmewahrscheinlichkeit der Teilnehmenden und Nichtteilnehmenden bestimmt. Aufgrund der Nichtteilnahme liegen üblicherweise wenige Informationen für die Nichtteilnehmenden vor. Entsprechend beläuft sich die Analyse auf die Informationen, die für alle Beobachtungseinheiten vorhanden sind, und hängt stark von der Datenverfügbarkeit ab. So verwendet z. B. das SOEP entweder aggregierte Regionaldaten für Erstbefragte oder Daten aus den Vorwellen bei Panelaustritten (Kroh, Kühne, Goebel & Preu, 2015) zur Prüfung einer selektiven Nicht-Teilnahme. Alternativ wurde im Rahmen der Querschnittbefragung "*Mobility in Cities* – SrV 2013" im Anschluss eine kurze Befragung einer Zufallsauswahl von Nichtteilnehmenden durchgeführt (Wittwer & Hubrich, 2015). Mit Hilfe der verfügbaren Informationen können dann die Teilnahmewahrscheinlichkeiten aller (Nicht-)Teilnehmer:innen bestimmt werden. Korreliert diese Wahrscheinlichkeit stark mit beobachteten Merkmalen der Stichprobe, kann dies auf *nonresponse bias* hindeuten: Personen mit bestimmten Ausprägungen dieser Merkmale sind eher dazu geneigt, freiwillig an Befragungen teilzunehmen. Dies würde dazu führen, dass diese Personengruppen überproportional oft oder selten in der Stichprobe vertreten sind. Somit wäre die Stichprobe nicht mehr repräsentativ. Endsprechend ist bei dieser Analyse die Annahme wichtig, dass die Teilnahmewahrscheinlichkeit auf beobachteten Informationen zurückzuführen ist und nicht von unbeobachteten Variablen abhängt (*Missing at Random*, Little & Rubin, 2002).

2.2 Gewichtungsmethoden

Ein Lösungsansatz ist die Verwendung von Gewichten. Das Ziel ist, mit einer Umgewichtung der einzelnen Beobachtungen die Repräsentativität der Stichprobe herzustellen. Hierzu empfehlen Kalton und Flores Cervantes (2003), die Gewichtung in drei Schritten vorzunehmen. Zuerst wird die per Design erzeugte über- und unterproportionale Vertretung der Teilgruppen korrigiert. Hierzu wird meist die Inverse der Ziehungswahrscheinlichkeit verwendet, die im Stichprobendesign festgelegt wurde (Kim & Kim, 2007). Im zweiten Schritt werden die Gewichte angepasst, um *unit nonresponse* zu berücksichtigen. Dabei können die Ergebnisse der Ausfallanalyse verwendet werden. Im dritten Schritt wird die Poststratifizierung durchgeführt, welche die erhaltenen Gewichte nochmals korrigiert, um die Stichprobe wieder der Grundgesamtheit anzupassen.

Es gibt verschiedene Methoden, um die Gewichtung durchzuführen. Die Wahl der geeigneten Methoden hängt vom Stichprobendesign und der Stichprobengröße ab. In einigen Fällen ist auch die Kombination von verschiedenen Methoden sinnvoll. Im Folgenden werden die Methoden *cell weighting*, *raking* und logistische Regression diskutiert. Dabei werden jeweils die Idee, die Berechnung der Gewichte, mögliche Einschränkungen sowie Anwendungsbeispiele aufgeführt. Kalton und Flores Cervantes (2003) führen eine ausführliche Diskussion inklusive vereinfachtem Rechenbeispiel auf.

Cell weighting: Beim *cell weighting* werden die Beobachtungen der Stichprobe und der Grundgesamtheit nach den Ausprägungen bestimmter kategorialer Merkmale gruppiert. Es soll dann für jede relevante Merkmalskombination dieselbe Gewichtung in der Stichprobe und Grundgesamtheit erreicht werden. So soll z. B. der Anteil verheirateter Frauen in der Stichprobe und Grundgesamtheit gleich sein. Je mehr Merkmale (und Merkmalsausprägungen) bei dieser Gewichtung berücksichtigt werden, desto mehr Zellen entstehen. Unter der Annahme, dass innerhalb jeder Zelle die *Missing at Random* (MAR) Annahme zutrifft, errechnet sich das Gewicht für alle Beobachtungen einer Zelle indem die Zahl der Beobachtungen innerhalb dieser Zelle von der Grundgesamtheit (G_{jk}) ins Verhältnis zur Stichprobe (S_{jk}) gesetzt wird.

Am Beispiel von zwei Merkmalen, ergibt sich der Hochrechnungsfaktor (HRF) w_{jk} für alle Beobachtungen der Zelle (j, k):

$$w_{jk} = \frac{G_{jk}}{S_{jk}} \quad . \tag{1}$$

Ein potentielles Problem des *cell weightings* ist die Möglichkeit, dass einzelne Zellen in der Stichprobe nicht besetzt bzw. vertreten sind. Dies kann vor allem dann auftreten, wenn es insgesamt sehr viele Zellen, also Merkmalskombinationen, gibt. Ähnlich kann dies dazu führen, dass nur sehr wenige Beobachtungen in einer Zelle vorhanden sind. In diesem Fall erhalten diese Beobachtungen große Gewichte, was zu einer erhöhten Varianz und verringerter Präzision der Stichprobenschätzer führen kann (Kalton & Flores Cervantes, 2003). Im ersten Fall können Merkmalsausprägungen zusammengefasst werden (Kalton & Flores Cervantes, 2003). Alternativ werden wie z. B. beim Gesundheits-Monitoring-Einheiten in Bayern irrelevante Merkmale mit Hilfe logistische Regressionen ausfindig gemacht und in der Gewichtung nicht berücksichtigt (Kass et al., 2021). Im zweiten Fall können die Gewichte auf einen festgelegten maximal Wert begrenzt werden, wie z. B. im IAB-Betriebspanel (Bachmann et al., 2020).

Raking: Alternativ wird beim *raking* nicht jede Merkmalskombination (also einzelne Zelle) berücksichtigt, sondern die Verteilung der einzelnen kategorialen Merkmale (also die Randsummen). Bei der Berechnung der Gewichte, werden die Beobachtungen zuerst auf die Randsummen eines ersten Merkmals hochgerechnet. Die der ersten Hochrechnung entsprechend gewichteten Stichprobenbeobachtungen werden dann auf Basis des nächsten Merkmals hochgerechnet. Das Verfahren verläuft analog für alle weiteren aus der Grundgesamtheit bekannten Merkmale weiter. Als iteratives Verfahren wird der Prozess für alle Merkmale so oft wiederholt, bis die Gewichte konvergieren.

Das folgende Beispiel zeigt die ersten Iterationen der Berechnung des HRF für Zelle (j, k) bei zwei Merkmalen mit J bzw. K Ausprägungen (in Anlehnung an die Notation von Battaglia, Izrael, Hoaglin & Frankel, 2004). Hierbei entspricht für die ungewichtete Stichprobe das ursprüngliche Gewicht $w_i = 1$ für alle Beobachtungen. Damit wird mit w_{jk} die Summe der Gewichte in Zelle (j, k) und mit w_{j+} und w_{+k} die Summe der Gewichte pro Zeile und Spalte bezeichnet. G_{j+} und G_{+k} entsprechen den Zeilen- und Spaltensummen der Grundgesamtheit. Aufgrund der iterativen Struktur dieses Verfahrens erhalten wir nach jeder Iteration ein Gewicht, wobei $w_{jk}^{(1)}$ dem Gewicht nach der ersten Iteration entspricht.

$$w_{jk}^{(0)} = w_{jk} \qquad j = 1, \ldots, J;\ k = 1, \ldots, K \tag{2}$$

$$w_{jk}^{(1)} = w_{jk}^{(0)} \left(\frac{G_{j+}}{w_{j+}^{(0)}} \right) \qquad \text{für alle } k \text{ innerhalb allen } j \tag{3}$$

$$w_{jk}^{(2)} = w_{jk}^{(1)} \left(\frac{G_{+k}}{w_{+k}^{(1)}} \right) \qquad \text{für alle } j \text{ innerhalb allen } k. \tag{4}$$

Es wird immer abwechselnd über die Zeile- und Spaltensummen umgewichtet bis die Gewichte konvergieren, so dass für Iteration s (wobei $s = 0, 1, \ldots$) gilt:

$$w_{jk}^{(2s+1)} = w_{jk}^{(2s)} \left(\frac{G_{j+}}{w_{j+}^{(2s)}} \right) \tag{5}$$

$$w_{jk}^{(2s+2)} = w_{jk}^{(2s+1)} \left(\frac{G_{+k}}{w_{+k}^{(2s+1)}} \right). \tag{6}$$

Bei der Anwendung von *raking* wird erneut angenommen, dass innerhalb jeder Zelle die MAR Annahme zutrifft. Zusätzlich wird angenommen, dass sich die Teilnahmewahrscheinlichkeit jeder Zelle als Produkt der jeweiligen Randwahrscheinlichkeiten ergibt. Falls diese (schärferen) Annahmen nicht zutreffen, kann potentiell eine neue Verzerrung entstehen. Im Vergleich dazu ist die Varianz der Gewichte beim *raking* auch bei vielen Zellen geringer als beim *cell weighting* (Kalton & Flores Cervantes, 2003). Infolgedessen bietet sich *cell weighting* an, wenn die Gesamtzahl an Zellen gering ist, während bei einer hohen Zahl an Zellen *raking* sinnvoller erscheint. Abschließend stellt *raking* eine gute Wahl dar, wenn in der Grundgesamtheit keine Information über die gemeinsame Merkmalsverteilung vorhanden ist, da in diesem Fall *cell weighting* nicht möglich ist.

Logistische Regression: Eine flexible Alternative ist die Berechnung von Gewichten mithilfe der individuellen Teilnahmewahrscheinlichkeiten, auch *propensity scores* genannt. Hier findet also die oben angesprochene Ausfallanalyse Anwendung, um für einen eventuellen *nonresponse bias* zu korrigieren. Basierend auf beobachteten Merkmalen für Teilnehmende und Nichtteilnehmende wird mit einer logistischen Regression geschätzt, wie hoch die Teilnahmewahrscheinlichkeit der einzelnen Beobachtungseinheiten ist. Die Gewichte sind dann definiert als die inverse Teilnahmewahrscheinlichkeit der jeweiligen Beobachtung *i*:

$$w_i = \frac{1}{\hat{\pi}_i} = \frac{e^{\widehat{logit}_i} + 1}{e^{\widehat{logit}_i}}, \tag{7}$$

wobei $\hat{\pi}_i$ die geschätzte Antwortwahrscheinlichkeit darstellt.

Die berechneten Gewichte können denen vom *raking* sehr ähnlich werden, sofern die unabhängigen Variablen in der logistischen Regression alle kategorial sind. Allerdings können die Gewichte im unteren Bereich abweichen, da das minimale Gewicht bei logistischen Regressionen im Gegensatz zum *raking* nicht kleiner werden kann als 1: Bei der maximalen Teilnahmewahrscheinlichkeit von 1 (also 100%) ergibt das kleinste Gewicht $w_i = \frac{1}{1} = 1$. Nichtsdestotrotz bietet die logistische Regression mehr Flexibilität als *raking* in der Berücksichtigung von Merkmalen, da bei der Regression auch stetige Merkmale und Interaktionsterme aufgenommen werden können. Demgegenüber steht der Nachteil eines höheren Anspruchs an die Datensammlung, da hier die jeweiligen Informationen auch für die Nichtteilnehmer:innen vorhanden sein müssen.

Kombinationen: Zunehmend finden auch Kombinationen der Methoden praktische Anwendung. So empfehlen Kalton und Flores Cervantes (2003) zum Beispiel, *cell weighting* für Zellen zu nutzen, die ausreichend viele Beobachtungen enthalten, während für Zellen mit wenige Beobachtungen dann eine andere Methode, zum Beispiel *raking*, hinzugezogen werden kann. Ein weiteres Beispiel ist das SOEP, für welches bei Wiederholungsstichproben zunächst eine längsschnittliche Ausfallanalyse mit logistischer Regression und im Anschluss Randanpassungen mit dem *raking* Verfahren durchgeführt werden (Kroh et al., 2015).

3 Datengrundlagen und Ausfallanalyse

In diesem Kapitel wird zunächst das ZäPP selbst vorgestellt, dessen Ziel die Schaffung einer Datengrundlage zur vertragszahnärztlichen Versorgung in Deutschland darstellt. Wie erläutert, werden die Ergebnisse im ZäPP in der Regel gewichtet ausgewiesen. Die Gewichtung berücksichtigt dabei die Organisationsform, die KZV-Zugehörigkeit sowie die Umsatzklasse der teilnehmenden Zahnarztpraxen. Diese drei Merkmale sowie die Herkunft der Daten wird in den folgenden Paragraphen näher erläutert. Zuletzt werden Art und Herkunft sozioökonomischer Kontrollvariablen vorgestellt, die dem Datensatz zum Zwecke einer Ausfallanalyse hinzugefügt werden.

3.1 Datengrundlagen

Zahnärzte-Praxis-Panel (ZäPP): Das ZäPP ist eine bundesweite Kostenstrukturerhebung in Vertragszahnarztpraxen, die vom Zentralinstitut für die kassenärztliche Versorgung in Deutschland (Zi) im Auftrag der KZBV durchgeführt wird. Es handelt sich um eine Wiederholungsbefragung von Praxen der kassenzahnärztlichen Versorgung zur Praxisstruktur, zu erbrachten zahnärztlichen Leistungen sowie den Praxisaufwendungen und -einnahmen. Praxen der kassenzahnärztlichen Versorgung umfassen Praxen der allgemeinen Stomatologie, der Kieferorthopädie sowie der Oral- und Mund-Kiefer-Gesichtschirurgie.

Das ZäPP ist als repetitive Querschnittserhebung mit jährlichen Erhebungswellen konzipiert. In jeder Erhebungswelle werden die beiden der Erhebung vorhergehenden Kalenderjahre (Berichtsjahre) abgefragt. Die Erhebungswelle 2020 deckt beispielsweise die Berichtsjahre 2018 und 2019 ab und die Erhebungswelle 2019 die Berichtsjahre 2017 und 2018. Somit wird jedes Berichtsjahr faktisch zweimal erhoben. Die Befragung startet jedes Jahr im September und endet im März des Folgejahres. Die Ergebnisse werden nach Prüfung und Auswertung der eingereichten Daten in Form eines Tabellenbandes jeweils im Juni an die KZBV übergeben.

Die Grundgesamtheit des ZäPP umfasst alle Einzelpraxen und BAGs in Deutschland, die kassenzahnärztliche Leistungen über die jeweils zuständige KZV abrechnen. Die Definition erfolgt auf Basis der Registereintragungen bei der jeweils zuständigen KZV. Reine Privatpraxen und Medizinische Versorgungszentren sowie ihnen gleichgestellte Einrichtungen gemäß den §§ 95 bzw. 311 des fünften Sozialgesetzbuches werden im ZäPP nicht berücksichtigt und sind nicht Teil der Grundgesamtheit.

Für die Erhebungen des ZäPP werden aus der Grundgesamtheit alle Praxen ausgewählt, die sich während des Berichtszeitraums nicht durch Organisationsformwechsel, Umzug oder ähnliches grundlegend verändert und über den vollständigen Berichtszeitraum kassenzahnärztliche Leistungen erbracht haben. Diese ausgewählten Praxen bilden die Auswahlgesamtheit und werden mit der Bitte um Teilnahme am ZäPP angeschrieben. In der Erhebungswelle 2020 handelte es sich dabei um ca. 38.000 Praxen. Ansprechpartner sind die Praxisinhaber. In den vergangenen Erhebungswellen konnten je Erhebungswelle zwischen 3.300 und 4.500 Teilnehmer gewonnen werden. Die Ausschöpfungsquote lag somit bei rund 10%. Die Teilnahme am ZäPP ist freiwillig. Eine Teilnahme- und Auskunftspflicht besteht nicht.

Jede angeschriebene Praxis erhält einen Fragebogen, der in drei Teile gegliedert ist. Der erste Teil umfasst allgemeine Angaben zur Praxis, u.a. zur Organisationsform, den Praxisräumen, Laborbetrieb, Personal und Arbeitszeiten. Im zweiten Teil werden Angaben zu den erbrachten zahnärztlichen Leistungen der gesetzlichen und privaten Krankenversicherung abgefragt. Der dritte Teil des Fragebogens beinhaltet Angaben zu den Einnahmen und Aufwendungen der Praxis auf Grundlage der steuerlichen Einnahmenüberschussrechnung. Die Angaben bzw. die Richtigkeit der Angaben zu den Finanzen der Praxis müssen von einem Steuerberater/einer Steuerberaterin oder einer ähnlich qualifizierten Person testiert werden. Jeder eingesendete Fragebogen wird auf Vollständigkeit und das Vorliegen eines Testats geprüft. Nur wenn beide Bedingungen erfüllt sind, werden die Fragebögen erfasst und in die weitere Verarbeitung übernommen.

Jede teilnehmende Praxis erhält für die Einsendung eines vollständig ausgefüllten und testierten Fragebogens eine finanzielle Aufwandspauschale sowie persönliche Feedbackberichte nach Abschluss der Erhebung, die einen Vergleich der Kennzahlen der Praxis mit den Werten der Praxen des jeweiligen KZV-Bereichs ermöglichen. Nach Erfassung der Angaben aus dem Fragebogen werden die gewonnenen Daten im Zi validiert.

Die validierten und geprüften Daten sind Grundlage der anschließenden Auswertungen. Die Ergebnisse werden an die KZBV und die jeweilige KZV übergeben. Die KZBV erhält darüber hinaus vom Zi den pseudonymisierten Analysedatensatz. Die Ergebnisse des ZäPP werden auf vielfältige Weise genutzt. Sie fließen in das Jahrbuch der KZBV ebenso ein wie in die Vergütungsverhandlungen mit den gesetzlichen Krankenkassen.

Da die Teilnahme am ZäPP freiwillig ist, ist eine Selbstselektion der teilnehmenden Praxenmöglich. Variierende Ausschöpfungsquoten der Teilnehmenden führen zu Abweichungen der Verteilung zwischen den am ZäPP teilnehmenden Praxen und den Praxen in der Grund- bzw. Auswahlgesamtheit im Hinblick auf wichtige Strukturmerkmale. Durch Gewichtung der Beobachtungen und entsprechende Hochrechnung kann einer möglichen Verzerrung der Ergebnisse entgegengewirkt werden. Daher erfolgen Analysen des Zahnärzte-Praxis-Panels in der Regel gewichtet.

Daten aus den Zahnarztregistern: Aus den Zahnarztregistern der jeweils zuständigen KZV liegen für alle Praxen der Auswahlgesamtheit die Angaben zur Organisationsform (Einzelpraxis oder BAG), zur Anzahl und zum Geschlecht der Praxisinhaber sowie zur jeweils zuständigen KZV vor. Diese Daten können dem Zi in anonymisierter Form zur Verfügung gestellt werden. Ergänzt werden diese Angaben um die Information, ob die Praxis am ZäPP teilgenommen hat oder nicht (Teilnehmerstatus). Zu Zwecken der *nonresponse*-Korrektur werden die Daten für Teilnehmende des ZäPP um das Teilnahme-Pseudonym ergänzt.

Umsatzklassenstatistiken: Für das ZäPP erstellen die KZV außerdem sogenannte Umsatzklassenstatistiken auf Grundlage ihrer Abrechnungsdaten. Diese Umsatzklassenstatistiken weisen separat für EP und BAG die Anzahl und den Gesamtumsatz aller Praxen einer Umsatzklasse in der jeweiligen KZV aus. Die Umsatzklassen sind in 25.000 EUR-Schritte unterteilt. Die Zuordnung einer Praxis zu einer Umsatzklasse ergibt sich aus ihren im entsprechenden Kalenderjahr über die KZV vereinnahmten Umsätzen.

Räumliche Daten: Im Rahmen dieses Papiers werden den Daten der Auswahlgesamtheit zusätzliche regionale sozioökonomische Daten zugespielt. Auf Postleitzahlebene kann das Merkmal Regionsdichte zugeordnet werden, das auf den Angaben zum Grad der Verstädterung gemäß Gemeindeverzeichnis des Statistischen Bundesamtes mit Stand zum 31.12.2017 basiert (Statistisches Bundesamt (Destatis), 2018).

Außerdem werden ausgewählte Merkmale aus den Indikatoren und Karten zur Raum- und Stadtentwicklung (INKAR) herangezogen (Bundesinstitut für Bau-, Stadt,- und Raumforschung, 2021). Diese Daten liegen auf Landkreisebene vor und enthalten eine Vielzahl von regionalstatistischen Informationen. Daraus wurden für diese Arbeit die folgenden Daten ausgewählt: Arbeitslosenquote, durchschnittliches Haushaltseinkommen, Ausländeranteil, Durchschnittsalter, Ärzte je 10.000 Einwohner, Wahlbeteiligung an der Bundestagswahl sowie die entsprechenden Anteile der Zweitstimmen für die Parteien CDU/CSU, SPD, Bündnis 90/Die Grünen, FDP, Die Linke und AfD. Die Merkmale aus der INKAR-Datenbank beziehen sich auf das Jahr 2017.

Abschließend wird das Merkmal Erreichbarkeit von Zahnärzten ergänzt, das die Median-Wegezeit in PKW-Minuten gemäß des Thünen-Erreichbarkeitsmodells auf Landkreisebene zum Stand September 2016 misst (Bundesministerium für Ernährung und Landwirtschaft, 2021).

Diese Raumdaten werden im nächsten Abschnitt für eine Ausfallanalyse verwendet, mit der ein *unit nonresponse* geprüft und gegebenenfalls korrigiert werden soll.

3.2 Ausfallanalyse

Die vorliegenden anonymisierten Praxismerkmale aus den Zahnarztregistern werden mit den Raumdaten gemäß Abschnitt 3.1 verknüpft und um einen Teilnahmeindikator ergänzt. Eine Teilnahme liegt vor, wenn die Praxis einen vollständigen und durch den Steuerberater bestätigten Fragebogen zur Erhebung 2020 des ZäPP eingereicht hat. Im Rahmen einer Ausfallanalyse können potenzielle Unterschiede zwischen Teilnehmenden und Nicht-Teilnehmenden aufgezeigt werden.

Empirische Strategie: In der Ausfallanalyse soll identifiziert werden, ob und falls ja welche Merkmale die Teilnahmebereitschaft beeinflussen. Hierzu wird die Teilnahme als abhängiges Merkmal und die auf individueller und Raumebene vorliegenden Informationen als unabhängige Merkmale in einem Logit-Modell betrachtet. Dabei werden in Tabelle 2 drei Modellspezifikationen präsentiert. Die erste Modellspezifikation (Spalte 1) berücksichtigt zunächst nur die Merkmale, die aus den Zahnarztregistern bekannt sind (Organisationsform, Geschlecht, KZV-Bereich und Aufwandspauschale in 100 Euro). Alternativ werden in der zweiten Spezifikation regionale Informationen betrachtet (Spalte 2). Zuletzt werden alle verfügbaren Informationen in die Regression aufgenommen (Spalte 3). In der Tabelle werden die geschätzten durchschnittlichen marginalen Effekte und ihre Standardfehler (in Klammern) dargestellt. Sie geben die Veränderung in der Teilnahmewahrscheinlichkeit wider, die mit einem Anstieg der korrespondieren Merkmale um eine Einheit einhergeht. Dabei kann ein positiver (negativer) und signifikanter marginaler Effekt als positiver (negativer) Einfluss auf die Teilnahme durch das entsprechende Merkmal interpretiert werden. Statistische Signifikanz auf dem 1%-, 5%- und 10%-Niveau wird mit Sternen (***/**/*) markiert. Liegt keine statistische Signifikanz vor, so liegt auch keine Evidenz für einen Zusammenhang vor.

Ergebnisse: In der erste Spalte zeigen sich regionale Unterschiede in der Teilnahmebereitschaft: Für Betriebe in Sachsen ist die Teilnahmebereitschaft 5,4 Prozentpunkte höher als im Referenz KVZ-Bereich Schleswig-Holstein. Mit einer durchschnittlichen Teilnahmebereitschaft von 6,7% in Schleswig-Holstein, entspricht dieser marginale Effekt einer erhöhten Bereitschaft von 80% in Sachsen (im Vergleich zu Schleswig-Holstein). Auch in anderen KVZ-Bereichen ist die Bereitschaft höher, z. B. in Rheinland-Pfalz und in Sachsen-Anhalt. Dies könnte beispielsweise auf eine höhere Akzeptanz der Umfrage in diesen KVZ-Bereichen zurückzuführen sein. Zusätzlich ist noch auffällig, dass zwischen BAG und EP Praxen kein Unterschied in der Bereitschaft nachgewiesen werden kann und auch kein Effekt von der Aufwandspauschale ersichtlich ist.

In der zweiten Spezifikation sind ebenfalls einzelne signifikante Effekte zu verzeichnen. Insbesondere zeigt sich eine sinkende Teilnahmebereitschaft mit steigendem regionalem Haushaltseinkommen und positive Ef-

fekte bei den Wahlergebnissen. Wichtig hierbei ist aber, dass eben diese starken Effekte verschwinden, sobald zusätzlich für die regionalen Unterschiede kontrolliert wird (Spalte 3). Dies deutet daraufhin, dass diese Effekte durch regionale Unterschiede getrieben werden, welche bereits im bisherigen Gewichtungsverfahren im ZäPP berücksichtigt werden (vergleiche Kapitel 4.1). Insgesamt muss auch festgehalten werden, dass die berechneten Pseudo-R^2 der drei Modelle auf einen geringen Erklärungsgehalt hinweisen.

Somit kann geschlussfolgert werden, dass in der Tat regionale Unterschiede vorliegen. Solche regional variierende Ausschöpfungsquoten wurden auch in vergangenen Erhebungswellen des ZäPP erkannt (für Details siehe Abschnitt 4.1). Davon abgesehen, liefert die Ausfallanalyse aber keine starke empirische Evidenz für das Vorliegen systematischer Unterschiede zwischen Teilnehmer:innen und Nicht-Teilnehmer:innen basierend auf den verfügbaren Merkmalen.

Diese Schlussfolgerung kann auch bei der Betrachtung der aus den Logit-Regression resultierenden *Propensity Scores* gezogen werden. Hierbei werden für alle Praxen basierend auf den Ergebnissen der Logit-Regressionen die Teilnahmewahrscheinlichkeit, dem sogenannten *Propensity Score*, geschätzt. Abbildung 1 zeigt die Verteilungen der Propensity Scores zwischen Teilnehmer:innen und Nicht-Teilnehmer:innen für die drei Modellspezifikationen. Es zeigt sich in allen drei Fällen eine klare Spiegelung der Teilnahmewahrscheinlichkeit zwischen den Teilnehmer:innen und Nicht-Teilnehmer:innen. Daher kann das Fazit gezogen werden, dass im ZäPP kein erkennbarer *nonresponse bias* vorliegt und eine entsprechende Korrektur nicht notwendig ist.

Zu einem ähnlichen Ergebnis kommen auch andere Panel-Erhebungen. Zum Beispiel wurde für das IAB- Betriebspanel die Ausfallanalyse für wiederholte Befragungen durchgeführt (Janik & Kohaut, 2009). In diesem Fall liegen weitreichende Informationen aus dem Vorjahr vor und es konnte zwischen unkontrollierbaren (z. B. die generelle ökonomische Situation der Betriebe) und kontrollierbaren Faktoren (z. B. die Erfahrung der befragenden Person) unterschieden werden. Janik und Kohaut (2009) kommen zu dem Schluss, dass die

Tabelle 2 Ausfallanalyse: ZäPP Teilnahme (Marginale Effekte)

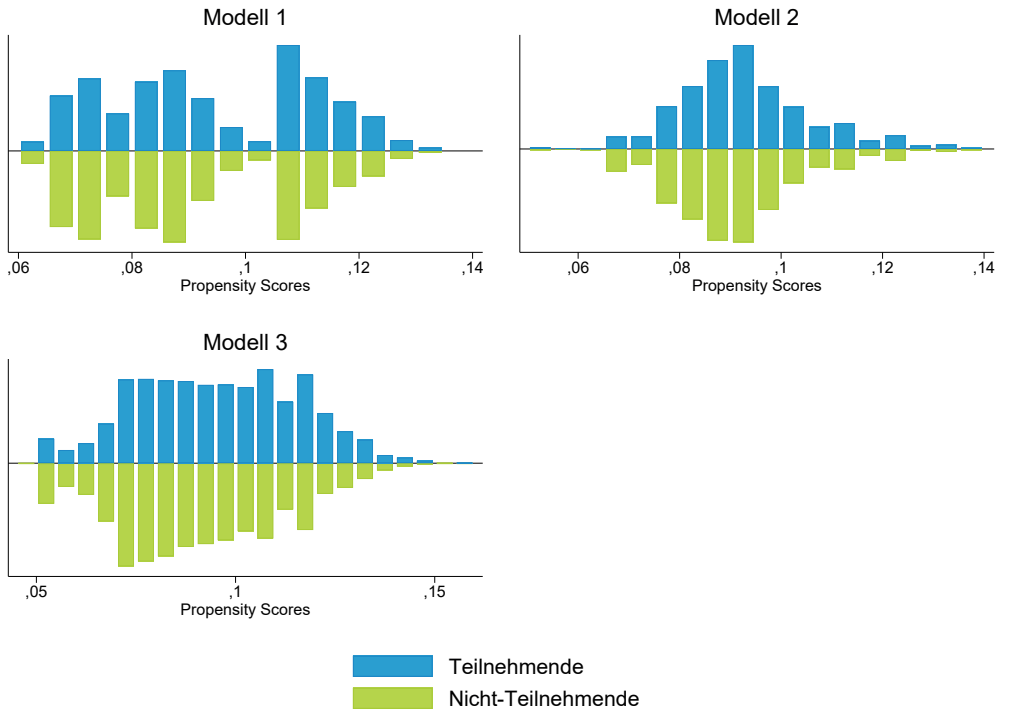
	Logit-Mod. (1)	Logit-Mod. (2)	Logit-Mod. (3)
Organisationsform (Ref. BAG) Einzelpraxis	0,0049 (0,0123)		0,0068 (0,0121)
Geschlecht (Ref. männlich) weiblich	0,0068** (0,0032)		0,0076** (0,0032)
KZV-Bereich (Ref. Schlesw.Holst.) Hamburg	0,0078 (0,0117)		0,0321** (0,0136)
Bremen	0,0230 (0,0204)		0,0475** (0,0230)
Niedersachsen	0,0142 (0,0092)		0,0143 (0,0088)
Westfalen-Lippe	0,0181* (0,0096)		0,0238** (0,0099)
Nordrhein	0,0250*** (0,0090)		0,0364*** (0,0103)
Hessen	0,0113 (0,0093)		0,0309*** (0,0099)
Rheinland-Pfalz	0,0497*** (0,0117)		0,0537*** (0,0121)
Baden-Württemberg	0,0216 (0,0266)		0,0407 (0,0270)
Bayern	0,0036 (0,0083)		0,0234* (0,0137)
Berlin	0,0251 (0,0273)		0,0684* (0,0373)
Saarland	0,0177 (0,0161)		0,0296 (0,0194)
Mecklenburg-Vorpommern	0,0227* (0,0128)		0,0418 (0,0260)

	Logit-Mod. (1)	Logit-Mod. (2)	Logit-Mod. (3)
Brandenburg	0,0311** (0,0151)		0,0501* (0,0269)
Sachsen-Anhalt	0,0360*** (0,0120)		0,0521* (0,0267)
Thüringen	0,0253** (0,0114)		0,0463* (0,0261)
Sachsen	0,0540*** (0,0107)		0,0820** (0,0329)
Aufwandspauschale in 100 EUR	0,0081 (0,0097)		0,0093 (0,0097)
Regionsdichte (Ref. dicht besiedelt) mittlere Besiedlungsdichte		0,0095* (0,0056)	0,0046 (0,0058)
gering besiedelt		0,0166** (0,0069)	0,0139* (0,0071)
Zahnarzt-Erreichbarkeit in Minuten		-0,0023 (0,0015)	-0,0008 (0,0019)
Haushaltseinkommen in 1.000 EUR		-0,0446*** (0,0139)	-0,0214 (0,0171)
Ausländeranteil in %		-0,0009 (0,0006)	-0,0023*** (0,0008)
Durchschnittsalter in Jahren		0,0002 (0,0017)	-0,0001 (0,0019)
Ärzte je 10.000 Einwohner		-0,0007 (0,0006)	0,0013* (0,0008)
Wahlbeteiligung in %		0,0001 (0,0008)	-0,0008 (0,0010)
Wahlergebnisse (Ref. sonst. Parteien) Anteil CDU/CSU		0,0022 (0,0016)	-0,0001 (0,0026)
Anteil SPD		0,0019 (0,0015)	0,0008 (0,0026)
Anteil B90/Grüne		0,0043*** (0,0016)	0,0005 (0,0029)
Anteil FDP		0,0032** (0,0013)	0,0007 (0,0030)
Anteil Linke		0,0024 (0,0017)	-0,0022 (0,0033)
Anteil AfD		0,0035* (0,0018)	-0,0005 (0,0030)
Pseudo R^2	0,006	0,003	0,008
Observations	36.089	36.089	36.089

Quellen: Zahnärzte-Praxis-Panel 2020, Statistisches Bundesamt, INKAR (Bundesinstitut für Bau-, Stadt-, und Raumforschung), Landatlas (Bundesministerium für Ernährung und Landwirtschaft), eigene Berechnungen. Anmerkungen: Die Tabelle präsentiert marginale Effekte der unabhängigen Variablen basierend auf Logit-Regressionen mit dem ZäPP-Teilnahme Dummy als abhängige Variable. Das Merkmal Regionsdichte liegt auf Postleitzahlebene vor. Die anderen raumbezogenen Merkmale von Zahnarzt-Erreichbarkeit bis Wahlergebnisse liegen auf Landkreisebene vor. Standardfehler sind in Klammern angegeben. * p < 0,1, ** p < 0,05, *** p < 0,01

meisten untersuchten Faktoren keinen Einfluss auf die Teilnahmebereitschaft haben, oder durch das Studiendesign bzw. Interview kontrolliert werden können. Daher wird auch im IAB-Betriebspanel keine zusätzliche Korrektur durchgeführt, die einen potenziellen *nonresponse bias* berücksichtigt. Entsprechend stellt die Schlussfolgerung der vorliegenden Studie kein Einzelphänomen dar.

Abbildung 1 **Propensity Scores aus der Ausfallanalyse**



Quelle: Zahnärzte-Praxis-Panel 2020, Statistisches Bundesamt, INKAR (Bundesinstitut für Bau-, Stadt-, und Raumforschung), Landatlas (Bundesministerium für Ernährung und Landwirtschaft), eigene Berechnungen. Anmerkungen: Die Abbildung zeigt die Propensity Scores der Teilnehmer:innen und Nicht-Teilnehmer:innen basierend auf den in Tabelle 2 durchgeführten Logit-Regressionen.

4 Gewichtung im Zahnärzte-Praxis-Panel

Das folgende Kapitel fokussiert auf die Gewichtung im ZäPP. Dabei wird zunächst das bisherige Gewichtungsverfahren, das zuletzt in der Erhebung des ZäPP 2020 zum Einsatz kam, detailliert vorgestellt. Die Hochrechnung erfolgte per *cell weighting* unter Einhaltung einer Mindestbesetzungszahl von sechs Beobachtungen je Zelle. Aufgrund der Feingliederung der Hochrechnungsrahmen kommt es dabei häufig zu Nichtbesetzungen und Unterschreitungen der Mindestbesetzungszahl in einzelnen Zellen. Dadurch müssen Zellen zusammengelegt werden. Die Zusammenlegung erfolgt dabei bestmöglich nach dem Prinzip der Gleichverteilung. Bisher wurde diese Gleichverteilung der Elemente auf die einzelnen Zellen in den Hochrechnungsrahmen, also in der Grundgesamtheit, verwirklicht. Da dieses Verfahren erst erfolgreich ist, wenn die Beobachtungen aus der Stichprobe unter Einhaltung der Mindestbesetzungszahl in die resultierenden Zellen gesetzt werden können, handelt es sich dabei um eine rekursive Methode.

Alternativ könnte die Zusammenlegung von Zellen unter dem Aspekt der Gleichverteilung auch direkt auf Basis der Stichprobenverteilung detailliert vorgestellt durchgeführt werden. Dieses Vorgehen hat praktische Vorteile: zum einen, dass Beobachtungen in den Randzellen bei der Gleichverteilung schneller berücksichtigt werden können, und zum anderen erfolgt die Prüfung der Mindestbesetzung nicht mehr rekursiv. Das Kriterium der Gleichverteilung wäre dann in den Zellen der Stichprobe verwirklicht und nicht wie bisher in den Zellen der Grundgesamtheit. Da die Auswirkungen einer Umstellung auf das alternative Verfahren zur Zusammenlegung von Zellen analytisch schwer zu beurteilen sind, erfolgt die Bewertung und der Vergleich zwischen dem alternativen und dem bisherigen Verfahren anhand einer Monte-Carlo-Simulation, deren Durchführung und Ergebnisse im Abschnitt 4.2 präsentiert werden.

Im dritten Unterabschnitt erfolgt ein Vergleich der verschiedenen konkret auf das ZäPP angewendeten Gewichtungsmethoden. Bei diesem Vergleich werden die Mittelwertschätzungen der KZV-Umsätze anschließend mit Hilfe der aus der Grundgesamtheit bekannten tatsächlichen Mittelwerte auf die Schätzgenauigkeit hin untersucht.

4.1 Bisheriges Gewichtungsverfahren

Das Gewichtungsverfahren im ZäPP wurde bei der ersten Erhebungswelle 2018 festgelegt und orientiert sich am Zi-Praxis-Panel (ZiPP), einer ähnlich angelegten Kostenstrukturerhebung in Praxen der kassenärztlichen Versorgung. Es handelt sich dabei um ein *cell weighting*-Verfahren, das seit der ersten Erhebungswelle geringfügig weiterentwickelt wurde. Im folgenden wird das Verfahren beschrieben, wie es zuletzt im ZäPP 2020 zum Einsatz kam.

Grundlage für die Gewichtung sind die in 3.1 vorgestellten Umsatzklassen der über die jeweilige KZV vereinnahmten Umsätze. Die Umsatzklassen sind in 25.000 EUR-Schritten klassiert und werden getrennt nach der Organisationsform (EP/ BAG) ausgegeben. Dementsprechend liegen die Daten zu Umsatzklassen, Organisationsform und KZV-Zugehörigkeit für die Grundgesamtheit klassiert-kombiniert vor.

Auf dieser Grundlage werden dann Gewichtungszellen je Organisationsform und KZV-Zugehörigkeit (Auswahlgesamtheit) gebildet, die ein oder mehrere Umsatzklassen umfassen. Um zu verhindern, dass einzelne Gewichtungsklassen ein zu großes Gewicht erhalten, ist festgelegt, dass jeder Gewichtungszelle mindestens 6 Beobachtungen aus der Stichprobe zugeordnet werden müssen. Deshalb wird zunächst (unter Berücksichtigung der Anzahl entsprechender Praxen in der Stichprobe und der dort besetzten Umsatzklassen) berechnet, wie viele Gewichtungszellen maximal gebildet werden können. Anschließend erfolgt die Erstellung dieser Gewichtungszellen nach dem Gleichbesetzungsprinzip aus der Grundgesamtheit. Das heißt, dass eine annähernd gleiche Besetzungszahl von Praxen aus der Auswahlgesamtheit je Gewichtungszelle angestrebt wird. Die gebildeten Gewichtungszellen werden dann mit Beobachtungen aus der Stichprobe besetzt und danach auf die Mindestbesetzung überprüft. Ist das Kriterium nicht erfüllt, wird die Anzahl der zu bildenden Gewichtungszellen um eins reduziert und die Gewichtungszellen werden nach gleichem Vorgehen neu gebildet. Dieses Verfahren läuft iterativ solange, bis in allen Klassen die Mindestbesetzung gewährleistet ist. Die Anzahl der Gewichtungszellen wird folglich unter den Annahmen der Gleich- und Mindestbesetzung maximiert, um eine differenziertere Hochrechnung auf die Praxen der Auswahlgesamtheit zu ermöglichen. Es ist anzumerken, dass eine vollständige Gleichbesetzung aufgrund der Einteilung in Honorarklassen nicht genau gewährleistet werden kann, sondern abhängig von den vorliegenden Daten so gleichmäßig wie möglich erfolgt.

Sollten in einem KZV-Bereich nicht ausreichend Praxen vorhanden sein, um je Organisationsform mindestens zwei Gewichtungszellen zu besetzen, werden KZV-Bereiche für die Gewichtung zusammengelegt, die eine geografische Nähe zueinander aufweisen und sich strukturell ähnlich sind. Nach Bildung der Gewichtungszellen erfolgt die Hochrechnung der Anzahl der Praxen aus der Stichprobe auf die Anzahl der Praxen in der Grundgesamtheit. In einem abschließenden Schritt werden die Gewichte normiert, damit die Summe der Gewichte der Anzahl der teilnehmenden Praxen entspricht. Die jeweiligen Gewichte können anschließend den entsprechenden Praxen angespielt und bei der Berechnung der Ergebnisse berücksichtigt werden.

Die Ergebnisse des ZäPP werden nicht nur auf Bundes- sondern auch auf KZV-Ebene ausgewiesen. Dementsprechend ist für die Daten auf KZV-Ebene ein eigenes Gewichtungsverfahren notwendig, bei dem der KZV-Bereich nicht berücksichtigt wird. Die Gewichtungszellen werden nach dem gleichen Verfahren nur je Organisationsform gebildet. Sollten bei der Gewichtung auf KZV-Ebene in einem KZV-Bereich nicht ausreichend Praxen vorhanden sein um je Organisationsform mindestens zwei Gewichtungszellen zu besetzen, wird im jeweiligen KZV-Bereich auf die Unterscheidung nach Organisationsform verzichtet. Die Gewichtung erfolgt dann lediglich nach den Umsatzklassen. Methodisch ist das Verfahren ansonsten identisch. Die jeweiligen Gewichte können anschließend den Praxen angespielt und bei der Berechnung der Ergebnisse berücksichtigt werden.

4.2 Simulation

Methode: Wie im vorigen Unterkapitel erläutert, werden Gewichtungszellen bei geringer Besetzung zusammengelegt. Hierbei wird bisher das Prinzip der Gleichbesetzung innerhalb der Grundgesamtheit unterstellt. Im Folgenden wird mit Hilfe einer Monte-Carlo-Simulation untersucht, ob alternativ das Prinzip der Gleichbesetzung innerhalb der Stichprobe angebracht wäre. Entsprechend bezeichnet Methode 1 (M1) die Zusam-

menlegung von Zellen unter Berücksichtigung der Gleichverteilung von Elementen in der Grundgesamtheit, während sich Methode 2 (M2) auf die Zusammenlegung unter Berücksichtigung der Gleichverteilung der Stichprobenbeobachtungen bezieht.

Neben dem Kriterium der Gleichverteilung werden Zellen im ZäPP unter Berücksichtigung einer Mindestbesetzungszahl je Zelle zusammengelegt. Die Höhe der Mindestbesetzungszahl hat Einfluss auf die Anzahl der resultierenden Gewichtungszellen. Die Simulation wird daher zusätzlich für eine variierende Anzahl zu bildender Zellen durchgeführt, um für einen Einfluss der Zellanzahl auf das Ergebnis zu kontrollieren.

Parameter: Der Rahmen der Simulation wird weitestgehend in Anlehnung an das ZäPP konstruiert. Dafür wird eine fiktive Grundgesamtheit mit 40.000 Elementen gebildet. Diese Größenordnung entspricht grob der ZäPP-Auswahlgesamtheit. Die Elemente sind Merkmalsträger für ein standardnormalverteiltes Merkmal. Die Wahrscheinlichkeit, in die Stichprobe gezogen zu werden, ist für jedes Element individuell und hängt linear steigend vom Merkmal mit Werten zwischen 3% und 15% ab. Die steigende Abhängigkeit vom Merkmal entspricht dem im ZäPP beobachteten stärkeren Rücklauf von Praxen mit höheren Umsätzen. Das Merkmal wird vorbereitend auf die später erfolgende Hochrechnung in 61 Klassen aufgeteilt, was der KZV-Umsatzklassenanzahl im ZäPP entspricht.

Durchführung: Vor der Stichprobenziehung wird der wahre Mittelwert des interessierenden Merkmals aus der Grundgesamtheit als Benchmark für die spätere Schätzung aus der Stichprobe gebildet. Auf Basis der berechneten Stichprobenziehungswahrscheinlichkeiten wird dann die Stichprobe gezogen. Die gewählten Stichprobenziehungswahrscheinlichkeitsgrenzen von 3% und 15% führen zu einer Stichprobengröße von etwa 9%, was in etwa der Ausschöpfungsquote im ZäPP 2020 entspricht. Aus der Wahrscheinlichkeitsfunktion der Stichprobenziehung resultiert eine von der Verteilung der Grundgesamtheit rechts verschobene Stichprobenverteilung. Der Mittelwert der ungewichteten Stichprobe ist deshalb nicht erwartungstreu.

Im Anschluss an die Stichprobenziehung erfolgt die Zusammenlegung der Merkmalsklassen nach den Methoden M1 und M2 bei einer steigenden Anzahl resultierender Gewichtungszellen von 2 bis 15. Auf Basis der resultierenden Gewichtungszellen erhalten die Beobachtungen der Stichprobe im Anschluss Hochrechnungsfaktoren, die je Zelle als Quotient der Anzahl der Elemente der Grundgesamtheit geteilt durch die Anzahl der Beobachtungen der Stichprobe definiert sind (*cell weighting*). Der entsprechend der Hochrechnung gewichtete Mittelwert aus der Stichprobe wird anschließend gespeichert und die Stichprobenziehung 500 mal wiederholt, um den Einfluss des Zufalls gering zu halten.

Für jede der beiden Methoden M1 und M2 und für jede Anzahl resultierender Gewichtungszellen werden 500 gewichetete Mittelwerte berechnet, die mit dem wahren Mittelwert der Grundgesamtheit verglichen werden. Je kleiner die Abweichung ist, desto besser ist die Schätzung. Um auch der Varianz der aus den Stichproben berechneten Mittelwerte gerecht zu werden, wird für jede der Methoden M1 und M2 und für jede Anzahl resultierender Gewichtungszellen der Mean Squared Error (MSE) als Summe der Varianz und der quadrierten durchschnittlichen Abweichung der geschätzten Mittelwerte berechnet. Je kleiner der MSE ist, desto besser ist der Schätzer.

Ergebnisse: Die Ergebnisse der Simulation sind geeignet, die beiden Verfahren miteinander zu vergleichen, da sie unter sonst gleichen Bedingungen durchgeführt werden.

In Abbildung 2 werden die resultierenden MSE in Abhängigkeit zur Anzahl der gebildeten Gewichtungszellen in einem Balkendiagramm dargestellt. Jeder Balken repräsentiert dabei das Ergebnis der 500 mal wiederholten Stichprobenziehung aus der fiktiven Grundgesamtheit jeweils nach Hochrechnung mit einer der beiden Methoden M1 und M2.

Grundsätzlich ist der Zusammenhang zwischen der Anzahl Gewichtungszellen und dem MSE bei beiden Verfahren hyperbolisch und sinkend mit steigender Anzahl gebildeter Zellen. Die Schätzung wird also im Hinblick auf Erwartungstreue und Streuung besser, je mehr Gewichtungszellen gebildet werden. Die Auswirkungen auf die Schätzgüte sind dabei besonders im Bereich weniger Gewichtungszellen spürbar. Der Sprung von zwei

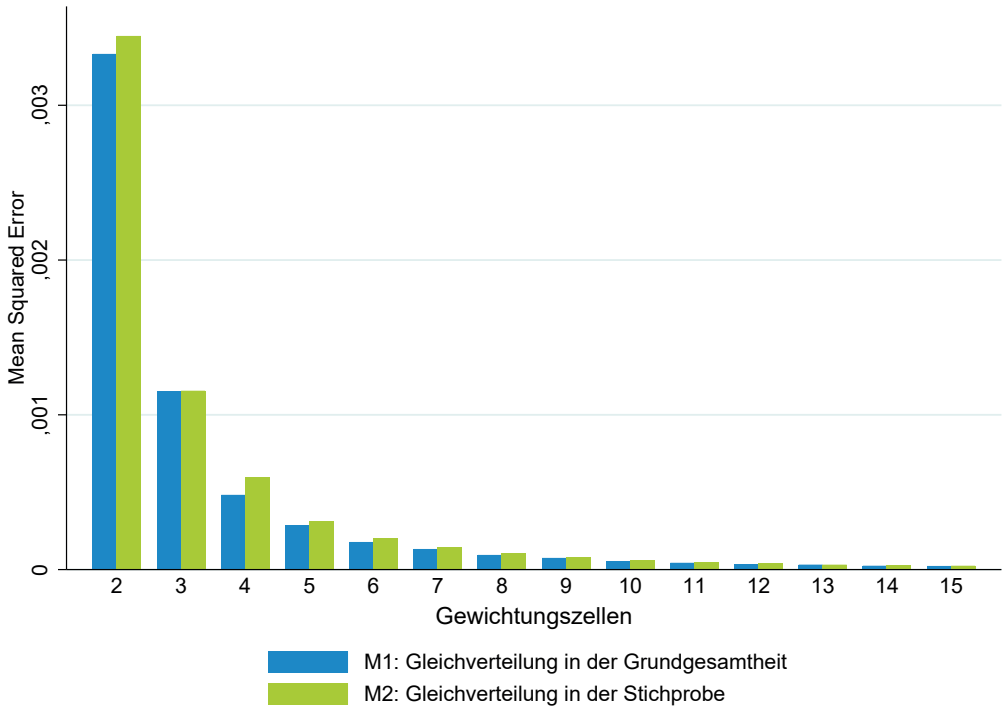
auf drei Gewichtungszellen verbessert die Schätzgüte vergleichsweise stärker als der Sprung von drei auf vier usw.

Der Vergleich der Methoden M1 und M2 zeigt, dass für jede Anzahl gebildeter Gewichtungszellen (mit Ausnahme von 13 Gewichtungszellen) der MSE bei Methode M1 kleiner als bei M2 ist.

Aufgrund des allgemeinen Ansatzes der Simulation könnte daher für die Hochrechnung im ZäPP geschlussfolgert werden, dass die Zusammenlegung von Zellen grundsätzlich weiterhin nach dem Prinzip der maximal möglichen Anzahl Gewichtungszellen sowie nach Methode M1 erfolgen sollte. Entscheidend für die Schätzgüte ist gemäß der Simulation jedoch weniger die Methode der Zusammenlegung von Zellen als vielmehr die Anzahl der gebildeten Gewichtungszellen. Daher ist es sinnvoll zu prüfen, wie viele Gewichtungszellen nach den Methoden M1 und M2 unter Berücksichtigung einer Mindestbesetzungszahl tatsächlich maximal gebildet werden können.

Anzahl Gewichtungszellen: Tabelle 3 zeigt die maximal mögliche Anzahl zu bildender Gewichtungszellen der kombinierten Merkmale KZV-Zugehörigkeit, KZV-Umsätze sowie Organisationsform für die Beobachtungen des ZäPP 2020 nach KZV-Bereichen und Organisationsform unter Berücksichtigung einer Mindestbesetzungszahl je Zelle von sechs Praxen. Für die EP können grundsätzlich mehr Gewichtungszellen gebildet werden als für BAG. Das liegt daran, dass EP unter allen Praxen einen Anteil von über 80% haben. In KZV-Bereichen mit vergleichsweise kleiner Anzahl Praxen wie Bremen kommt es mitunter nicht zu einer erfolgreichen Zellbildung, wodurch die Organisationsform in ebendiesen KZV-Bereichen nicht bei der Gewichtung berücksichtigt werden kann. Die Zusammenlegung von Zellen bei EP nach der Methode M1 führt in 5 KZV-Bereichen zu mehr Gewichtungszellen als mit der Methode M2. Umgekehrt führt die Zusammenlegung von Zellen bei EP mit der Methode M2 in 10 KZV-Bereichen zu mehr Zellen. In den KZV-Bereichen Nordrhein und Bremen führen beide Methoden zur selben Anzahl Zellen bei den EP. Klarer wird das Ergebnis bei Betrachtung der BAG. In keinem KZV-Bereich können mit der Methode M1 mehr Gewichtungszellen bei den BAG gebildet werden als mit Methode M2. In den KZV-Bereichen Schleswig-Holstein, Hamburg und Thüringen führt die Verwendung

Abbildung 2 **Mean Squared Errors nach der Simulation**



Quelle: Eigene Simulationsberechnung.
Anmerkungen: Die Abbildung zeigt den berechneten Mean Squared Error der gewichteten Mittelwertschätzung eines standardnormalverteilten Merkmals nach 500-fach wiederholter Stichprobenziehung aus einer fiktiven Grundgesamtheit, jeweils nach Zusammenlegung von Gewichtungszellen nach dem Kriterium der Gleichverteilung der Elemente in der Grundgesamtheit (M1) bzw. in der Stichprobe (M2) mit einer variierenden Anzahl resultierender Gewichtungszellen von 2 bis 15.

der Methode M1 gar nicht erst zu einer erfolgreichen Zellzusammenlegung, während mit der Methode M2 jeweils zwei Gewichtungszellen erfolgreich besetzt werden können.

Da die Mindestbesetzungszahl von sechs Praxen zu keiner zielführenden Gewichtungszellenbildung in sechs KZV-Bereichen führt, kann eine Herabsetzung der Mindestfallzahl auf drei Praxen je Gewichtungszelle analog zum ärztlichen Pendant des ZäPP, dem Zi-Praxis-Panel (ZiPP), das ebenfalls vom Zi durchgeführt wird, diskutiert werden. Tabelle 4 stellt die Ergebnisse von Tabelle 3 für eine Mindestbesetzungszahl von 3 Praxen dar. Im Gegensatz zu den Ergebnissen in Tabelle 3 führt die Herabsetzung der Mindestbesetzungszahl dazu, dass in 10 KZV-Bereichen die Zellzusammenlegung bei EP nach der Methode M1 zu mehr Gewichtungszellen führt als mit der Methode M2. Bei den BAG zeigt sich im Vergleich zur Mindestbesetzungszahl von 6 Praxen, dass die Zusammenlegung von Zellen in den Bereichen Schleswig-Holstein, Hamburg und Thüringen auch nach der Methode M1 zu einer erfolgreichen Zellzusammenlegung führt.

Tabelle 3 Anzahl gebildeter Gewichtungszellen mit Mindestbesetzung von sechs Praxen

	EP		BAG	
KZV-Bereich	M1	M2	M1	M2
Schleswig-Holstein	**6**	5	0	**2**
Hamburg	4	**5**	0	**2**
Bremen	2	2	0	0
Niedersachsen	15	**16**	4	**8**
Westfalen-Lippe	11	**13**	3	**5**
Nordrhein	18	18	5	**8**
Hessen	9	**12**	3	**5**
Rheinland-Pfalz	11	**13**	2	**4**
Baden-Württemberg	**20**	19	8	**11**
Bayern	**20**	19	7	**8**
Berlin	13	**14**	3	**4**
Saarland	3	**4**	0	0
Mecklenburg-Vorpommern	**8**	6	0	0
Brandenburg	8	**10**	2	2
Sachsen-Anhalt	7	**10**	2	2
Thüringen	10	**11**	0	**2**
Sachsen	**15**	14	5	5

Quellen: Zahnärzte-Praxis-Panel 2020.
Anmerkungen: Die Tabelle präsentiert die Anzahl gebildeter Gewichtungszellen bei einer Mindestbesetzung von 6 Praxen pro Zelle getrennt nach Einzelpraxen (EP) und Berufsausübungsgemeinschaften (BAG) jeweils unter der Annahme der Gleichbesetzung innerhalb der Grundgesamtheit (M1) und innerhalb der Stichprobe (M2).

Fazit: Aus den Ergebnissen der Simulation kann im Zusammenhang mit den Gegebenheiten des ZäPP geschlussfolgert werden, dass die Zusammenlegung von Zellen nach der Methode M2 sinnvoll ist, wenn kleine Beobachtungszahlen die Konstruktion von Gewichtungszellen erschweren. Für das ZäPP bedeutet das vor allem auch vor dem Hintergrund sinkender Ausschöpfungsquoten, dass sich eine Verbesserung der Schätzergebnisse durch die Umstellung auf die Methode 2 sowie eine an das ZiPP angeglichene Anpassung der Mindestbesetzungszahl ergeben könnte. Im folgenden Abschnitt werden daher die Methoden M1 und M2 sowie Mindestbesetzungen von sechs und drei Praxen je Zelle für die in Abschnitt 2.2 präsentierten Gewichtungsverfahren *cell weighting* und *raking* konkret auf das ZäPP 2020 angewendet und die Ergebnisse miteinander verglichen.

Tabelle 4 Anzahl gebildeter Gewichtungszellen mit Mindestbesetzung von drei Praxen

	EP		BAG	
KZV-Bereich	M1	M2	M1	M2
Schleswig-Holstein	8	**14**	2	**4**
Hamburg	7	**9**	2	**4**
Bremen	4	**5**	0	**2**
Niedersachsen	**19**	17	4	**12**
Westfalen-Lippe	18	**20**	7	**10**
Nordrhein	**22**	18	10	**16**
Hessen	16	16	6	6
Rheinland-Pfalz	**17**	16	5	**6**
Baden-Württemberg	**23**	19	11	**17**
Bayern	**21**	19	10	**14**
Berlin	**17**	15	4	**8**
Saarland	**6**	5	0	0
Mecklenburg-Vorpommern	**12**	10	0	**2**
Brandenburg	12	**15**	4	**5**
Sachsen-Anhalt	**16**	15	2	**4**
Thüringen	12	**15**	3	**4**
Sachsen	**15**	14	7	**9**

Quellen: Zahnärzte-Praxis-Panel 2020.
Anmerkungen: Die Tabelle präsentiert die Anzahl gebildeter Gewichtungszellen bei einer Mindestbesetzung von 3 Praxen pro Zelle getrennt nach Einzelpraxen (EP) und Berufsausübungsgemeinschaften (BAG) jeweils unter der Annahme der Gleichbesetzung innerhalb der Grundgesamtheit (M1) und innerhalb der Stichprobe (M2).

4.3 Vergleich der Gewichtungsmethoden

Hierfür werden die geschätzten Mittelwerte der KZV-Umsätze zum Berichtsjahr 2019 mit den aus der Grundgesamtheit bekannten Mittelwerte verglichen. Je kleiner die Differenz zum Mittelwert der Grundgesamtheit ist, desto besser ist die Schätzung.

Vorgehen: Die Gewichtungsmethoden *cell weighting* und *raking* werden jeweils nach einer Zellzusammenlegung von Gewichtungszellen nach dem Kriterium der Gleichbesetzung in der Grundgesamtheit (M1) bzw. in der Stichprobe (M2) durchgeführt. Zudem erfolgen die Berechnungen für variierende Mindestbesetzungszahlen je Gewichtungszelle von sechs und drei Beobachtungen. Eine kleinere Mindestbesetzungszahl bedeutet eine potenziell größere Anzahl Gewichtungszellen, was gemäß der Simulation in Abschnitt 4.2 die Präzision der Schätzung erhöht. Die Setzung der Mindestbesetzungszahl auf weniger als drei Beobachtungen ist aus Gründen der Ausreißerempfindlichkeit nicht zu empfehlen und wird daher nicht in Betracht gezogen.

Die Kehrwerte der in Abschnitt 3.2 berechneten Propensity Scores aus den drei betrachteten Modellspezifikationen werden auch als Gewichte berücksichtigt und in den Vergleich der Gewichtungsmethoden aufgenommen. Zuletzt dient die Berechnung der ungewichteten Ergebnisse als weiterer Benchmark für den Erfolg der Gewichtungsmethode.
Tabelle 5 stellt die relative Abweichung vom Mittelwert der Grundgesamtheit in % als normiertes Gütemaß dar. Die Ergebnisse sind getrennt auf Bundesebene und auf KZV-Ebene dargestellt, da die Gewichtung auf Bundesebene mit dem Merkmal KZV-Zugehörigkeit ein weiteres Merkmal berücksichtigt. Außerdem erfolgt eine Unterscheidung zwischen allen Praxen, EP und BAG. Die Ergebnisse auf KZV-Ebene werden als Durchschnitt über alle KZVen zusammengefasst. Die Ergebnisse der jeweils am besten abschneidenden Verfahren sind fett markiert.

Ergebnisse auf Bundesebene: Auf Bundesebene ist die Schätzung mit *raking* nach Methode M1 am präzisesten über alle Praxen. Die Abweichung von den bekannten KZV-Umsätzen der Grundgesamtheit beträgt lediglich 0,54%. Beim *raking* auf Bundesebene hat die Variation in der Mindestbesetzungszahl je Zelle keinen Einfluss auf das Ergebnis. Das liegt daran, dass die Mindestbesetzungszahl beim *raking* weniger Einfluss als

beim *cell weighting* hat, da vor allem bei den KZV-Umsatzklassen keine weitere Unterscheidung zwischen KZV-Bereichen und der Organisationsform stattfindet. Dadurch sind die meisten Gewichtungszellen bereits von vornherein mit einer ausreichend großen Anzahl Praxen besetzt.

Der ungewichtete Mittelwert für alle Beobachtungen ist mit einer Abweichung von 0,84%bereits verhältnismäßig genau, was auf eine Überschätzung der KZV-Umsätze der Einzelpraxen und eine Unterschätzung der KZV-Umsätze der BAG in der ungewichteten Stichprobe zurückzuführen ist. Daher sind hier die relativen Abweichungen auf Subgruppenebene verhältnismäßig hoch und die Gewichtungsmethoden *cell weighting* und *raking* liefern allesamt deutlich präzisere Schätzungen. Das *raking* nach Methode M1 führt mit einer Abweichung von nur 0,01% in den Einzelpraxen und einer Abweichung von 1,84% in den BAG wieder jeweils zu den präzisesten Ergebnissen. Besonders bei der Schätzung der KZV-Umsätze der BAG ist das *cell weighting* auf Bundesebene sowie das *raking* nach Methode M2 deutlich unpräziser mit einer Verschlechterung der Präzision bis zu etwa 3,3 Prozentpunkten.

Die Gewichtung auf Basis der Ausfallanalyse (Logit-Modelle) führt in allen Fällen zu einem den ungewichteten Ergebnissen ähnlichen und in der Regel etwas unpräziseren Ergebnis. Dies bestätigt die Schlussfolgerung in Abschnitt 3.2, nach der keine Korrektur für einen potenziellen *nonresponse bias* auf Basis der beobachteten Merkmale erfolgen sollte. Eine weitere mögliche Erklärung ist, dass die beobachteten erklärenden Merkmale der Ausfallanalyse unkorreliert mit dem Merkmal KZV-Umsätze sind.

Tabelle 5 Relative Abweichung der KZV-Umsätze 2019 in %

	Bundesebene			KZV-Ebene		
Methode	Alle	EP	BAG	Alle	EP	BAG
Cell Weighting (M1, MBZ = 3)	1,25	0,12	3,95	2,55	1,84	7,18
Cell Weighting (M1, MBZ = 6)	1,64	0,25	4,95	2,76	2,34	8,91
Cell Weighting (M2, MBZ = 3)	0,87	0,03	2,88	2,31	**1,55**	**6,93**
Cell Weighting (M2, MBZ = 6)	1,47	0,08	5,15	2,40	1,88	7,46
Raking (M1, MBZ = 3)	**0,54**	**0,01**	**1,84**	2,52	2,96	9,42
Raking (M1, MBZ = 6)	**0,54**	**0,01**	**1,84**	2,76	2,69	9,74
Raking (M2, MBZ = 3)	1,22	0,16	4,50	**1,94**	2,59	8,29
Raking (M2, MBZ = 6)	1,22	0,16	4,50	2,45	2,60	9,55
Logit-Modell 1	0,80	4,16	10,25	4,66	6,93	11,75
Logit-Modell 2	0,92	3,42	10,93	4,65	6,66	11,82
Logit-Modell 3	1,32	3,62	10,78	4,74	6,78	11,62
Ungewichtet	0,84	3,55	10,66	4,48	6,69	11,82

Quellen: Zahnärzte-Praxis-Panel 2020, Umsatzklassenstatistiken der KZVen, eigene Berechnungen.
Anmerkungen: Die Tabelle zeigt die relativen Abweichung zwischen den geschätzten (gewichteten) Mittelwerten und dem bekannten Mittelwert der Grundgesamtheit des Merkmals KZV-Umsätze im Jahr 2019. In jeder Zeile wird angegeben, welches Gewichtungsverfahren verwendet wurde. Beim cell weighting und beim raking bezieht sich die Angabe in den Klammern auf die Methode der Zellzusammenlegung (M1: Gleichverteilung von Elementen in der Grundgesamtheit, M2: Gleichverteilung von Beobachtungen in der Stichprobe) sowie auf die definierte Mindestbesetzungszahl je Gewichtungszelle (MBZ). Die Nonresponse Modelle beziehen sich auf die Logit-Regressionen der Aufsfallanalyse in Tabelle 2. Die Bestimmung der gewichteten Mittelwerte und den korrespondierenden relativen Abweichungen werden zunächst auf Bundesebene (für alle Praxen, nur für Einzelpraxen (EP) und für Berufsausübungsgemeinschaften (BAG)) durchgeführt. Alle Abweichungen werden analog als durchschnittliche Abweichung auf KZV-Ebene dargestellt. Fettgedruckt sind die minimalen Abweichungen je Spalte.

Ergebnisse auf KZV-Ebene: Auf KZV-Ebene unterscheidet sich das beste Ergebnis nach Gruppen. Während *raking* nach Methode M2 über alle Beobachtungen mit einer Abweichung von 1,94% am präzisesten ist, liefert *cell weighting* nach Methode M2 bei einer Mindestbesetzungszahl von drei Praxen mit Abweichungen von 1,55% in den Einzelpraxen und 6,93% in den BAG die genauesten Schätzungen. Ein weiteres Muster ist auf KZV-Ebene klar erkennbar: Die Schätzungen werden genauer durch die Zellzusammenlegung nach Methode M2 im Vergleich zur Methode M1. Eine mögliche Erklärung ist, dass die Fallzahlen in vielen KZV-Bereichen auf Zellebene viel kleiner als auf Bundesebene sind und die Methode M2 bei kleiner Fallzahl mit größerer Wahrscheinlichkeit eine erfolgreiche Verteilung der Beobachtungen auf neue Gewichtungszellen verwirklicht als die Methode M1.

Schlussfolgerung: Allgemein bestätigen die Ergebnisse, dass eine Herabsetzung der Mindestbesetzungszahl (und einer damit steigenden Anzahl Gewichtungszellen) bei gleichbleibender Gewichtungsmethode

und gleichbleibender Methode der Zellzusammenlegung zu einer erhöhten Genauigkeit führt. Zum Beispiel beträgt die relative Abweichung beim *cell weighting* nach Methode M1 auf Bundesebene 1,25% bei einer Mindestbesetzungszahl von drei Praxen im Vergleich zu einer Abweichung von 1,64% bei einer Mindestbesetzungszahl von sechs Praxen. Wie aus der Simulation abgeleitet, wird durch diesen Vorteil eine Zunahme der Standardfehler in Kauf genommen. Tabelle 6, in der die relativen Standardfehler analog zur Darstellung in Tabelle 5 präsentiert werden, spiegelt genau diesen Trade-Off wider. Die relativen Standardfehler sind bei den ungewichteten Berechnungen, zum Beispiel 2,21% bei den BAG auf Bundesebene, sowie bei den Gewichtungen auf Basis der Ausfallanalyse niedriger als beim *cell weighting* und *raking*, deren relative Standardfehler bei den BAG auf Bundesebene zwischen 2,54% und 3,25% liegen. Demgegenüber stehen die Verbesserungen in der Schätzgenauigkeit beim *cell weighting* und *raking* mit Abweichungen zwischen 1,84% und 5,15% im Vergleich zu den Abweichungen von über 10% bei den ungewichteten sowie nach den gewichteten Ergebnissen auf Basis der Ausfallanalyse. Die Verbesserung durch die Gewichtung mittels *cell weighting* und *raking* kompensiert daher die Steigung der Standardfehler.

Tabelle 6 Relative Standardfehler der KZV-Umsatz-Schätzungen 2019 in %

	Bundesebene			KZV-Ebene		
Methode	Alle	EP	BAG	Alle	EP	BAG
Cell Weighting (M1, MBZ = 3)	1,43	1,31	2,75	5,75	5,47	9,82
Cell Weighting (M1, MBZ = 6)	1,33	1,21	2,55	5,23	5,12	8,98
Cell Weighting (M2, MBZ = 3)	1,43	1,23	2,87	5,72	5,12	10,02
Cell Weighting (M2, MBZ = 6)	1,33	1,22	2,54	5,24	5,06	9,04
Raking (M1, MBZ = 3)	1,56	1,34	3,25	5,88	5,63	10,09
Raking (M1, MBZ = 6)	1,56	1,34	3,25	5,60	5,61	9,92
Raking (M2, MBZ = 3)	1,38	1,30	2,62	5,98	5,66	10,50
Raking (M2, MBZ = 6)	1,38	1,30	2,62	5,67	5,58	10,19
Logit-Modell 1	1,04	1,02	**2,20**	**4,33**	4,29	8,57
Logit-Modell 2	1,04	1,00	2,22	4,37	4,29	8,57
Logit-Modell 3	1,05	1,03	2,21	4,40	4,36	8,65
Ungewichtet	**1,03**	**0,98**	2,21	4,34	**4,26**	8,53

Quellen: Zahnärzte-Praxis-Panel 2020, Umsatzklassenstatistiken der KZVen, eigene Berechnungen.
Anmerkungen: Die Tabelle zeigt die relativen Standardfehler der geschätzten (gewichteten) Mittelwerte für das Merkmal KZV-Umsätze im Jahr 2019. In jeder Zeile wird angegeben, welches Gewichtungsverfahren verwendet wurde. Beim cell weighting und beim raking bezieht sich die Angabe in den Klammern auf die Methode der Zellzusammenlegung (M1: Gleichverteilung von Elementen in der Grundgesamtheit, M2: Gleichverteilung von Beobachtungen in der Stichprobe) sowie auf die definierte Mindestbesetzungszahl je Gewichtungszelle (MBZ). Die Nonresponse Modelle beziehen sich auf die Logit-Regressionen der Aufsfallanalyse in Tabelle 2. Die relativen Standardfehler werden zunächst auf Bundesebene (für alle Praxen, nur für Einzelpraxen (EP) und für Berufsausübungsgemeinschaften (BAG)) ermittelt und analog als durchschnittlicher Standardfehler auf KZV-Ebene dargestellt. Fettgedruckt sind die minimalen relativen Standardfehler je Spalte.

Anwendung auf weitere Berichtsjahre: Zur Überprüfung der Ergebnisse wird der Vergleich der Gewichtungsmethoden mit den Daten der Berichtsjahre 2018 und 2017 wiederholt. Die Ergebnisse werden in den Tabellen A.1 und A.2 im Anhang präsentiert. Insbesondere für das Berichtsjahr 2018 bestätigt sich, dass auf Bundesebene *raking* nach Methode M1 sowie auf KZV-Ebene *cell weighting* nach Methode M2 zu sehr präzisen Ergebnissen führen. Die Ergebnisse für das Berichtsjahr 2017 weisen kein so eindeutiges Muster auf wie die Ergebnisse für 2018 und 2019. Allerdings liegt für das Berichtsjahr 2017 eine mit mehr als 4.500 Beobachtungen um über 25% größere Stichprobe vor als in den Folgejahren, was einen Einfluss auf die Gewichtung erwarten lässt.

Zusammenfassung: Zusammenfassend kann festgehalten werden, dass sowohl die mit *raking* als auch mit *cell weighting* gewichteten Ergebnisse unabhängig von der Methode (M1/M2) und unabhängig von der Mindestbesetzungszahl deutlich präziser sind als die ungewichteten Ergebnisse oder die Ergebnisse auf Grundlage der *Nonresponse* Modelle. Es zeigt sich auch, dass die Ergebnisse auf Bundes- und KZV-Ebene mit verschiedenen Verfahren gewichtet werden sollten, um die jeweils präzisesten Ergebnisse zu erhalten. Dabei ist sowohl die Methode (M1/M2) als auch die Mindestbesetzungszahl je Gewichtungszelle von Bedeutung und muss bei der Bewertung der Verfahren berücksichtigt werden.

5 Fazit

Das ZäPP hat als Kostenstrukturerhebung in Praxen der kassenzahnärztlichen Versorgung das Ziel, Praxisstrukturen, zahnärztliche Leistungen und wirtschaftliche Situation der Praxen möglichst präzise abzubilden. Da die Teilnahme am ZäPP freiwillig ist, ist eine Selbstselektion der Teilnehmenden möglich, die zu einer Verzerrung der Ergebnisse führen könnte. Aus diesem Grund erfolgen die Auswertungen im ZäPP in der Regel gewichtet. Die Gewichtung weist dabei einige Besonderheiten auf. Aus der Grundgesamtheit sind drei Merkmale bekannt: Die Honorarklassen, die KZV-Zugehörigkeit und die Organisationsform der Praxen (EP/BAG). Diese Angaben liegen kombiniert-klassiert für die drei Merkmale vor. Aufgrund der fein gegliederten Honorarklassen ergibt sich daher für das ZäPP eine große Anzahl von Gewichtungszellen, die aufgrund von Besetzungsproblemen auf Zellebene zusammengelegt werden. Die Zusammenlegung erfolgt dabei unter Einhaltung einer Mindestbesetzungszahl und unter dem Kriterium der Gleichverteilung von Elementen der Grundgesamtheit auf die resultierenden Gewichtungszellen. Abschließend muss berücksichtigt werden, dass die Ergebnisse des ZäPP sowohl auf Bundesebene als auch auf KZV-Ebene ausgegeben werden.

Ziel dieses Papiers war es, das Gewichtungsverfahren im ZäPP vor dem Hintergrund rückläufiger Teilnahmezahlen zu bewerten und zu prüfen, ob Anpassungen der Gewichtung zu präziseren Ergebnissen führen. Dazu wurden zunächst die Gründe für die Verzerrung von Stichproben erläutert und drei gängige Gewichtungsmethoden vorgestellt, das *cell weighting*, das *raking* sowie eine logistische Regression auf Basis einer Ausfallanalyse. Bei der Ausfallanalyse wurde untersucht, ob beobachtete Merkmale aus der Auswahlgesamtheit sowie zugeordnete raumbezogene Informationen Einfluss auf die Teilnahmebereitschaft einer Praxis haben. Die Ergebnisse weisen abgesehen von regionalen Unterschieden nicht auf eine bedeutende Selbstselektion hin, sodass eine entsprechende *nonresponse*-Korrektur als nicht notwendig erachtet wird.

Die Zusammenlegung von Gewichtungszellen wurde hinsichtlich der Kriterien der Mindestbesetzungszahl und der Gleichverteilung im Rahmen einer Simulationsberechnung, die in Anlehnung an das ZäPP konstruiert wurde, neu bewertet. Analog zum Kriterium der Mindestbesetzung wurde der Einfluss einer steigenden Anzahl von Gewichtungszellen auf die Schätzgüte untersucht. Alternativ zum Kriterium der Gleichverteilung von Elementen in der Grundgesamtheit (M1) wurde der Einfluss einer Gleichverteilung von Beobachtungen der Stichprobe (M2) auf die Gewichtungszellen untersucht. Die Simulation zeigte, dass die Schätzgüte mit steigender Anzahl Gewichtungszellen bei beiden Methoden M1 und M2 besser wird und dass die Methode M1 bei gegebener Anzahl Gewichtungszellen eine leicht bessere Schätzgüte aufweist. Nichtsdestotrotz konnte eine allgemeine Empfehlung für die Methode M1 nicht abgeleitet werden, weil die Zusammenlegung von Zellen nach den Verfahren M1 und M2 bei gegebenen Stichproben zu unterschiedlich vielen Gewichtungszellen führen können. Anhand des ZäPP wurde gezeigt, dass die Methode M2 bei kleinen Fallzahlen tendenziell zu einer höheren Anzahl Gewichtungszellen führt.

Die Ergebnisse der Simulation wurden dem abschließenden Methodenvergleich zugrunde gelegt, bei dem die Ergebnisse des ZäPP unter Verwendung der vorgestellten Gewichtungsmethoden (*cell weighting*, *raking*, logit) berechnet und die Mittelwerte der KZV-Umsätze mit dem bekannten Mittelwert der Grundgesamtheit verglichen wurden. Die Methoden *cell weighting* und *raking* führten unabhängig von der Methode (M1/M2) und der Mindestbesetzungszahl zu präzisen Ergebnissen.

Die Herabsetzung der Mindestbesetzungszahl führte bei gleichbleibenden Gewichtungsmethoden generell zu einer präziseren Schätzung, während die Standardfehler nur vergleichsweise schwach anstiegen. Vor dem Hintergrund der gesunkenen Ausschöpfungsquoten und der damit verbundenen Besetzungsproblematik einzelner Gewichtungszellen kann die Herabsetzung der Mindestbesetzungszahl von bisher sechs auf drei Praxen je Gewichtungszelle für zukünftige Auswertungen im ZäPP empfohlen werden.

Darüber hinaus zeigten sich im Detail gewisse Tendenzen in den Schätzergebnissen: Das *raking* nach Zusammenlegung von Zellen nach der Methode M1 führte bei den Vergleichen auf Bundesebene zu präziseren Schätzungen, während auf KZV-Ebene das *cell weighting* nach Zusammenlegung von Zellen nach der Methode M2 präziser war. Da bei der Gewichtung auf Bundesebene mit der KZV-Zugehörigkeit ein weiteres Gewichtungsmerkmal berücksichtigt wird, unterscheiden sich die Gewichtungsparameter auf Bundes- und

KZV-Ebene. Wie dargelegt bietet sich das *cell weighting* bei einer geringen Anzahl von Gewichtungszellen an, während das *raking* Vorteile bei einer hohen Anzahl von Gewichtungszellen mit sich bringt.

Aus diesen Gründen wird empfohlen, für die Gewichtung der ZäPP-Ergebnisse auf Bundes- und KZV-Ebene unterschiedliche Gewichtungsverfahren zu verwenden. Für die Gewichtung der ZäPP-Ergebnisse auf Bundesebene empfiehlt sich das *raking* nach Zusammenlegung von Zellen nach der Methode M1 und auf KZV-Ebene das *cell weighting* nach Zusammenlegung von Zellen nach der Methode M2.

Literaturverzeichnis

Bachmann, S., Tschersich, N., Ellguth, P., Kohaut, S. & Baier, E. (2020, Februar). Methoden- und Feldbericht zum IAB-Betriebspanel. FDZ-Methodenreport.

Battaglia, M. P., Izrael, D., Hoaglin, D. C. & Frankel, M. R. (2004). Tips and tricks for *raking* survey data (aka sample balancing). Abt Associates, 1, 4740–4744.

Blumenstiel, J. E. & Gummer, T. (2015). Prävention, Korrektur oder beides? In *Nonresponse bias* (S. 13–44). Springer.

Bundesinstitut für Bau-, Stadt,- und Raumforschung. (2021). INKAR - Indikatoren und Karten zur Raum- und Stadtentwicklung. Zugriff auf https://www.inkar.de/ (Zugriff: 12. Oktober 2021)

Bundesministerium für Ernährung und Landwirtschaft. (2021). Infoportal Zukunft.Land, Landatlas. Zugriff auf https://www.landatlas.de (Zugriff: 12. Oktober 2021)

Goebel, J., Grabka, M. M., Liebig, S., Kroh, M., Richter, D., Schröder, C. & Schupp, J. (2019). The German Socio-Economic Panel (SOEP). Jahrbücher für Nationalökonomie und Statistik, 239 (2), 345–360.

Groves, R. M. (2006). Nonresponse rates and nonresponse bias in household surveys. Public opinion quarterly, 70 (5), 646–675.

Groves, R. M. & Peytcheva, E. (2008). The impact of nonresponse rates on nonresponse bias: a meta-analysis. Public opinion quarterly, 72 (2), 167–189.

Janik, F. & Kohaut, S. (2009, Juli). Why don't they answer? - unit non-response in the IAB Establishment Panel. FDZ-Methodenreport.

Kalton, G. & Flores Cervantes, I. (2003, 01). Weighting methods. Journal of Official Statistics, 19.

Kass, B., Kutzora, S., Weinberger, A., Nennstiel, U., Heißenhuber, A., Herr, C. & Heinze, S. (2021). Poststratification as a suitable approach to generalize findings of two crosssectional studies along the bavarian compulsory school entrance examination: An exemplary poststratified analysis for asthma, hay fever and wheezing. International Journal of Hygiene and Environmental Health, 234.

Kim, J. K. & Kim, J. J. (2007). Nonresponse weighting adjustment using estimated response probability. Canadian Journal of Statistics, 35 (4), 501–514.

Kroh, M., Kühne, S., Goebel, J. & Preu, F. (2015). The 2013 IAB-SOEP Migration Sample (M1): Sampling design and weighting adjustment. SOEP Survey Papers (271).

Little, R. & Rubin, D. (2002). Statistical analysis with missing data. Wiley.

Statistisches Bundesamt (Destatis). (2018). Gemeindeverzeichnis - alle politisch selbständigen Gemeinden (mit Gemeindeverband) in Deutschland nach Fläche, Bevölkerung, Bevölkerungsdichte und der Postleitzahl des Verwaltungssitzes der Gemeinde. Ergänzt um die geografischen Mittelpunktkoordinaten, Reisegebiete und Grad der Verstädterung. Jahresausgabe 31.12.2017. Zugriff auf https://www.landatlas.dehttps://www.destatis.de/DE/Themen/Laender-Regionen/Regionales/Gemeindeverzeichnis/Administrativ/Archiv/GVAuszugJ/31122017_Auszug_GV.html (Zugriff: 12. Oktober 2021)

Wittwer, R. & Hubrich, S. (2015). Nonresponse in household surveys: A survey of nonrespondents from the repeated cross-sectional study "Mobility in Cities – SrV" in Germany. Transportation Research Procedia, 11, 66–84.

A. Anhang

Tabelle A.1 Relative Abweichung der KZV-Umsätze 2018 in %

	Bundesebene			KZV-Ebene		
Methode	Alle	EP	BAG	Alle	EP	BAG
Cell Weighting (M1, MBZ = 3)	2,93	0,60	7,30	3,86	1,59	9,33
Cell Weighting (M1, MBZ = 6)	2,68	0,36	7,01	3,52	1,36	10,06
Cell Weighting (M2, MBZ = 3)	2,08	**0,22**	5,28	**2,65**	**1,06**	**7,95**
Cell Weighting (M2, MBZ = 6)	2,88	0,49	7,40	3,58	1,07	9,54
Raking (M1, MBZ = 3)	1,47	0,58	**2,29**	3,57	2,20	9,83
Raking (M1, MBZ = 6)	1,47	0,58	**2,29**	3,65	1,56	10,35
Raking (M2, MBZ = 3)	1,74	0,55	3,29	2,90	1,64	9,39
Raking (M2, MBZ = 6)	1,74	0,55	3,29	3,08	1,89	9,39
Ungewichtet	**0,88**	3,45	10,63	5,87	4,94	13,16

Quellen: Zahnärzte-Praxis-Panel 2019, Umsatzklassenstatistiken der KZVen.
Anmerkungen: Die Tabelle zeigt die relativen Abweichung zwischen den geschätzten (gewichteten) Mittelwerten und dem bekannten Mittelwert der Grundgesamtheit des Merkmals KZV-Umsätze im Jahr 2018. In jeder Zeile wird angegeben, welches Gewichtungsverfahren verwendet wurde. Beim cell weighting und beim raking bezieht sich die Angabe in den Klammern auf die Methode der Zellzusammenlegung (M1: Gleichverteilung von Elementen in der Grundgesamtheit, M2: Gleichverteilung von Beobachtungen in der Stichprobe) sowie auf die definierte Mindestbesetzungszahl je Gewichtungszelle (MBZ). Die Bestimmung der gewichteten Mittelwerte und den korrespondierenden relativen Abweichungen werden zunächst auf Bundesebene (für alle Praxen, nur für Einzelpraxen (EP) und für Berufsausübungsgemeinschaften (BAG)) durchgeführt. Alle Abweichungen werden analog als durchschnittliche Abweichung auf KZV-Ebene dargestellt. Fettgedruckt sind die minimalen Abweichungen je Spalte.

Tabelle A.2 Relative Abweichung der KZV-Umsätze 2017 in %

	Bundesebene			KZV-Ebene		
Methode	Alle	EP	BAG	Alle	EP	BAG
Cell Weighting (M1, MBZ = 3)	0,96	0,30	2,51	2,08	1,61	5,20
Cell Weighting (M1, MBZ = 6)	0,78	0,19	2,16	1,81	1,76	5,57
Cell Weighting (M2, MBZ = 3)	**0,26**	0,25	**1,45**	2,05	1,69	4,42
Cell Weighting (M2, MBZ = 6)	0,64	0,27	2,77	2,28	**1,37**	6,16
Raking (M1, MBZ = 3)	0,62	**0,08**	2,25	**1,74**	1,98	4,62
Raking (M1, MBZ = 6)	0,62	**0,08**	2,25	1,85	1,75	5,96
Raking (M2, MBZ = 3)	0,49	0,17	2,02	1,76	1,87	5,97
Raking (M2, MBZ = 6)	0,49	0,17	2,02	1,75	1,78	5,80
Ungewichtet	1,24	2,44	10,75	5,72	4,96	12,16

Quellen: Zahnärzte-Praxis-Panel 2018, Umsatzklassenstatistiken der KZVen.
Anmerkungen: Die Tabelle zeigt die relativen Abweichung zwischen den geschätzten (gewichteten) Mittelwerten und dem bekannten Mittelwert der Grundgesamtheit des Merkmals KZV-Umsätze im Jahr 2017. In jeder Zeile wird angegeben, welches Gewichtungsverfahren verwendet wurde. Beim cell weighting und beim raking bezieht sich die Angabe in den Klammern auf die Methode der Zellzusammenlegung (M1: Gleichverteilung von Elementen in der Grundgesamtheit, M2: Gleichverteilung von Beobachtungen in der Stichprobe) sowie auf die definierte Mindestbesetzungszahl je Gewichtungszelle (MBZ). Die Bestimmung der gewichteten Mittelwerte und den korrespondierenden relativen Abweichungen werden zunächst auf Bundesebene (für alle Praxen, nur für Einzelpraxen (EP) und für Berufsausübungsgemeinschaften (BAG)) durchgeführt. Alle Abweichungen werden analog als durchschnittliche Abweichung auf KZV-Ebene dargestellt. Fettgedruckt sind die minimalen Abweichungen je Spalte.