Fachhochschule Nordwestschweiz
Hochschule für Wirtschaft

Gutachten zur Eignung der Rollenden Kostenstudie für die Taxpunktwertberechnung im Kanton Wallis

zuhanden der Ärztekasse Genossenschaft

Olten, 18. Juli 2023

Impressum

Bibliographische Angaben

Titel: Gutachten zur Eignung der Rollenden Kostenstudie für die Taxpunktwertberechnung im Kanton Wallis
Autoren: Schoch, T., Hulliger, B. und Thees, O.
Auftraggeber: Ärztekasse Genossenschaft
Ort: Olten
Datum: 18. Juli 2023

Projektteam

Tobias Schoch
Beat Hulliger
Oscar Thees

Kontakt

Prof. Dr. Tobias Schoch
Fachhochschule Nordwestschweiz
Hochschule für Wirtschaft, Institut ICC

Riggenbachstrasse 16
CH-4600 Olten, Schweiz

E-Mail: tobias.schoch@fhnw.ch
Tel. (direkt): +41 (0)62 957 21 02

Anmerkung

Der Bericht gibt die Auffassung des Projektteams wieder, die nicht notwendigerweise mit derjenigen des Auftraggebers bzw. der Auftraggeberin oder der Begleitorgane übereinstimmen muss. Für den Inhalt ist allein der Auftragnehmer / die Auftragnehmerin verantwortlich.

Inhaltsverzeichnis

Abkürzungsverzeichnis

AS	Amtliche Sammlung
BFS	Bundesamt für Statistik
BVGer	Bundesverwaltungsgericht
ESS	Europäisches Statistisches System
KVG	Krankenversicherungsgesetz (Bundesgesetz über die Krankenversicherung)
MAS	Strukturdaten der Arztpraxen und ambulanten Zentren (Erhebung)
OECD	Organisation for Economic Co-operation and Development
OKP	Obligatorischen Krankenpflegeversicherung
RoKo	Rollende Kostenstudie
SMVS	Société Médicale du Valais
TPW	Taxpunktwert

1 Einleitung

Ausgangslage

Der TARMED ist ein schweizweit einheitlicher Tarif für die Vergütung ambulanter Leistungen nach dem Bundesgesetz über die Krankenversicherung (KVG), dem Bundesgesetz über die Unfallversicherung (UVG), dem Bundesgesetz über Invalidenversicherung (IVG) und dem Bundesgesetz über die Militärversicherung (MVG). Die Vergütung der ambulant erbrachten Leistungen bemisst sich an der Anzahl der Taxpunkte (TP) multipliziert mit dem Taxpunktwert (TPW). Für die nach dem KVG abgerechneten Leistungen wird der TPW in jedem Kanton zwischen den Versicherern und den kantonalen Ärztegesellschaften ausgehandelt und im Tarifvertrag über die Vergütung ambulanter Leistungen der obligatorischen Krankenpflegeversicherung vereinbart. Scheitern die Verhandlungen, muss der Kanton den Taxpunktwert festsetzen.

Im Kanton Wallis sind die Verhandlungen gescheitert. Der Staatsrat setzte am 28. November 2018 den Taxpunktwert für Arztpraxen auf 0.84 CHF fest. Die Krankenversicherer sowie die Walliser Ärztegesellschaft (Société Médicale du Valais, SMVS) legten dagegen jeweils eigenständig Beschwerde ein. Das Bundesverwaltungsgericht (BVGer) hat im Mai 2022 über die Beschwerden zum Taxpunktwert für ambulante medizinische Leistungen entschieden und den Fall an den Staatsrat zurückgewiesen.[1]

Gegenstand

Die Ärztekasse Genossenschaft und die SMVS hatten den TPW auf Grundlage der Rollenden Kostenstudie (RoKo) berechnet und dem Staatsrat vorgelegt. Die RoKo ist eine von den kantonalen Ärztegesellschaften und der Ärztekasse schweizweit bei den selbständig praktizierenden Ärztinnen und Ärzten durchgeführte empirische Datenerhebung.

Der Staatsrat des Kantons Wallis und das BVGer haben die «Repräsentativität» der RoKo in Frage gestellt und den berechneten TPW zurückgewiesen.

Die Ärztekasse hat am 2. Februar 2023 der Fachhochschule Nordwestschweiz FHNW – Hochschule für Wirtschaft, vertreten durch Prof. Dr. Tobias Schoch und Prof. Dr. Beat Hulliger (unter Mitarbeit von Oscar Thees), den Auftrag erteilt, die Eignung der RoKo für die Berechnung des TPW im Kanton Wallis hinsichtlich «Repräsentativität» zu prüfen und darüber Bericht zu erstatten. Die Erkenntnisse der Prüfung sind im vorliegenden Gutachten dokumentiert.

Aufbau

Das Gutachten gliedert sich in drei Teile. Zunächst werden in Kapitel 2 die Grundlagen zum Begriff der «Repräsentativität» diskutiert und ein Qualitätsrahmen (engl. quality framework) vorgestellt, der zur Beurteilung der Qualität einer Erhebung herangezogen werden kann. In Kapitel 3 wird die RoKo als Erhebungsinstrument charakterisiert hinsichtlich ihrer Eignung für

[1] Urteil des BVGer C-7338/2018.

die Berechnung des TPW empirisch untersucht. Die Schlussfolgerungen des Gutachtens sind in Kapitel 4 festgehalten.

Anmerkung

Das Gutachten wird in Unabhängigkeit erstattet. Es nennt alle verwendeten Quellen und bezeichnet gegebenenfalls bestehende Unsicherheiten bei der Bewertung bestimmter Fragen. Wie üblich kann mit der Abgabe des Gutachtens nicht die Zusicherung verbunden sein, dass politische Behörden, Verwaltungsstellen oder Gerichtsbehörden bei der Beurteilung der entsprechenden Frage zu denjenigen Auffassungen gelangen, welche im vorliegenden Gutachten als zutreffend bezeichnet werden.

2 Grundlagen

2.1 Gütekriterien der empirischen Forschung

Die empirische Forschung orientiert sich an **Gütekriterien**. Diese bilden die Grundlage für eine angemessene Bewertung der Forschungsergebnisse. Gütekriterien werden für einzelne Messinstrumente und auch das Forschungsdesign insgesamt formuliert. Für Messinstrumente werden die Gütekriterien Objektivität, Reliabilität und Validität eingefordert. Über die zentrale Bedeutung dieser Kriterien, insbesondere deren Anspruch auf Allgemeingültigkeit, besteht in der empirischen Forschung ein breiter Konsens.[2]

Für Stichprobenerhebungen wird häufig – zusätzlich zu den oben genannten Gütekriterien – eingefordert, dass die Stichprobe «repräsentativ» zu sein habe. Im Zusammenhang mit der Festsetzung des TPW, nimmt das Bundesverwaltungsgericht (BVGer) in seinem Urteil C-7338/2018 vom 20. Mai 2022 diese Position betreffend die RoKo ein. Dabei greift das BVGer unter Bezugnahme auf das Buch von BENESCH[3] auf ein **metaphorisches Verständnis** von «Repräsentativität» zurück, wonach «Repräsentativität» dann gegeben sei, wenn die «Stichprobe ein verkleinertes, aber sonst wirklichkeitsgetreues Abbild der Grundgesamtheit» darstelle.[4] Nach welchen Kriterien zu beurteilen ist, ob eine wirklichkeitsgetreue Abbildung vorliegt, lässt das BVGer allerdings offen.

2.2 Der Begriff «Repräsentativität» und sinnvolle Alternativen

Entgegen einer weit verbreiteten Meinung sind «Repräsentativität» und «repräsentative Stichprobe» **keine Begriffe aus der Stichprobentheorie** (oder der Statistik im Allgemeinen). Der Begriff der «Repräsentativität» ist **kein quantifizierbares Merkmal** einer Stichprobe.[5] In der Forschungsliteratur wird der Begriff (bis auf wenige Ausnahmen) nicht verwendet.

Das im Urteil C-7338/2018 des BVGer zitierte Buch von BENESCH ist nicht der Forschungsliteratur zuzurechnen. Es handelt sich um ein Lehrbuch, das sich an Laien richtet. Aus diesem Grund hat BENESCH, wie er in der Einleitung des Buches schreibt, den «Formalisierungsgrad

[2] SCHNELL/HILL/ESSER, Methoden der empirischen Sozialforschung, 11., überarb. Aufl., 2018, S. 128 ff; ebenso DIEKMANN, Empirische Sozialforschung. Grundlagen, Methoden, Anwendungen, 14. Aufl., S. 230 ff.

[3] BENESCH, Schlüsselkonzepte zur Statistik: Die wichtigsten Methoden, Verteilungen, Tests anschaulich erklärt, 2013, S. 12.

[4] Urteil des BVGer C-7338/2018, S. 28 f: «[…] il importe toutefois que l'échantillon en question soit représentatif, à savoir qu'il permette de tirer des conclusions aussi exactes et sûres que possible sur l'ensemble de la « population ». C'est le cas lorsque le relevé partiel contient dans le mêmes proportions les caractéristiques du groupe visé, dont il représente une image certes réduite, sinon fidèle à la réalité (Thomas BENESCH, Schlüsselkonzepte zur Statistik, 2013, p. 9-12)».

[5] Für die deutschsprachige Literatur siehe bspw. DIEKMANN, a.a.O., S. 430; ebenso SCHNELL/HILL/ESSER, a.a.O., S. 297. Für die französischsprachige Literatur siehe bspw. TILLÉ, Théorie des Sondages: Échantillonnage et Estimation en Populations Finies, 2. Aufl., 2019, S. 74.

[...] so gering wie möglich gehalten».[6] Als Folge davon sind die von BENESCH vorgeschlagenen Begriffe und Definitionen zur Stichprobentheorie unpräzise – mitunter bis zur Unkenntlichkeit verzerrt – und genügen nicht wissenschaftlichen Ansprüchen.[7]

Alltags- und Verwaltungssprache

In der Alltags- und Verwaltungssprache zeichnet sich der Wortgebrauch von «Repräsentativität» im Zusammenhang mit Erhebungen durch ein hohes Mass an **Ambiguität** aus. Der mediale Diskurs über «repräsentative Stichproben» wird insbesondere von Markt- und Meinungsforschungsinstituten dominiert, die ihren Produkten die **Aura der Wissenschaftlichkeit** verleihen wollen («pseudo-scientific glamour»[8]). Der Begriff ist allerdings kaum mehr als «schmückendes und vorwiegend inhaltsleeres Attribut»[9].

Das Wort (bzw. der Begriff) «Repräsentativität» besitzt im öffentlichen Diskurs den Charakter eines Plastikwortes. **Plastikwörter** sind nach PÖRKSEN Wörter (oder Begriffe), die einen «diffusen und inhaltsarmen Universalitätsanspruch erheben»[10]. Beispiele sind «Information», «Struktur» oder «Prozess». Plastikwörter sind im Alltagsgebrauch so vage, dass man sie verwenden kann, ohne irgendetwas Substantielles zu sagen. Da Plastikwörter keine präzise Idee oder Vorstellung ausdrücken, können sie von fast allen akzeptiert werden, auch wenn die zugrunde liegenden Vorstellungen der Parteien weit auseinander liegen. Rein äusserlich sind Plastikwörter mit wissenschaftlichen Fachbegriffen verwandt, besitzen aber nicht deren präzise definierte Bedeutung. Sie sind nach und nach in die Alltags- und Verwaltungssprache eingegangen und erfreuen sich dort grosser Beliebtheit, da sie die Sprechenden mit der Aura des Expertentums umgeben und ihnen (vermeintliche) Deutungsmacht am Diskurs verleihen. Anstelle einer klaren Begriffsdefinition übertragen Plastikwörter eine «Autorität der Wissenschaftlichkeit in die Umgangssprache».[11] Der Gebrauch des Plastikwortes «signalisiert Wissenschaftlichkeit und bringt zum Schweigen».[12] Dazu gehört insbesondere der weit verbreitete autoritäre Reflex, einer Erhebung mit dem Worten «repräsentativ!» oder «nicht

---

6 BENESCH, a.a.O., S. V (Einleitung).

7 BENESCH, a.a.O., S. 145 gibt eine Definition von «Repräsentativität» an, die sich nur auf die (Stichproben-) Inklusionswahrscheinlichkeiten abstützt. Damit unterschlägt er, dass der grösste Teil der potenziellen Verzerrungen auf Teilnahme- oder Antwortverweigerungen zurückzuführen sind. Diese Definition von «Repräsentativität» greift viel zu kurz. Ferner ist seine Definition derart einschränkend formuliert, dass ihr nur ein einziges der gängigen einfachen Stichprobenverfahren genügt (nämlich die einfache Zufallsstichprobe mit Zurücklegen). Deutlich schwerer wiegt, dass seine Definition alle komplexen Stichprobenverfahren, wie. z. B. geschichtete oder geklumpte Stichprobenziehung und mehrstufige Ziehungsverfahren, ausschliesst und diesen Verfahren damit die «Repräsentativität», verstanden als Zulässigkeit eines Rückschlusses auf die Grundgesamtheit, per se abspricht.

8 KRUSKAL/MOSTELLER, Representative Sampling, I: Non-Scientific Literature, in: International Statistical Review 47, S. 13; siehe auch DIEKMANN, a.a.O., S. 430 f.

9 SCHNELL/HILL/ESSER, a.a.O., S. 296.

10 PÖRKSEN, Plastikwörter: Die Sprache einer internationalen Diktatur, 7. Aufl., S. 119 f.

11 PÖRKSEN, a.a.O., S. 121.

12 PÖRKSEN, a.a.O., S. 29.

repräsentativ!» die Glaubwürdigkeit zuzuerkennen oder abzusprechen, ohne dafür Argumente vorzubringen.

Begriffsauslegung

Aus den Arbeiten von KRUSKAL/MOSTELLER lassen sich zwei Begriffstypen zu «Repräsentativität» synthetisieren:[13]

- Repräsentativität als **Eigenschaftsbegriff** zur Charakterisierung der Stichprobe (sample), z. B. «miniature of the population»;
- Repräsentativität als **Erklärungsbegriff** zur Ziehung der Stichprobe (sampling); d. h. eine Erklärung, wie die Stichprobe generiert wurde, z. B. «representative sampling as permitting good estimation».

Der Eigenschaftsbegriff setzt an zu hoher Stelle an, weil er die realisierte Stichprobe an sich zu charakterisieren sucht, anstatt auf einzelne Messinstrumente und Teilaspekte zu fokussieren. Eine Stichprobe kann sehr wohl in einem Teilaspekt nicht «repräsentativ» sein, aber in vielen anderen Aspekten die Grundgesamtheit korrekt repräsentieren. Denkt man diesen Gedankengang konsequent zu Ende, so wird schnell klar, dass «Repräsentativität» als Eigenschaftsbegriff einer Erhebung oder Stichprobe kein sinnvoll definierter Begriff ist. Ein einfaches logisches Argument zeigt, dass eine Stichprobe (a fortiori) nicht repräsentativ für alle Merkmalsverteilungen einer Grundgesamtheit sein kann, da es in der Grundgesamtheit Kombinationen von Merkmalen geben kann, die in der Stichprobe (die ja nur eine Teilmenge der Grundgesamtheit darstellt) schlicht und einfach nicht vorhanden sind.[14]

Repräsentativität als Erklärungsbegriff aufzufassen, der den Ziehungsprozess einer Zufallsstichprobe erklärt, ist nicht abwegig. Allerdings ist diese Begriffsbildung für sich genommen nicht erhellend. Darauf wies bereits W. E. Deming mit der Aussage hin: «[w]hat we select are not representative samples, but probability samples»[15]. Es stellt sich nun zurecht die Frage, was der Begriff des «representative samplings» an begrifflichem Präzisionsgewinn leistet, das nicht bereits durch die Definition der Wahrscheinlichkeitsauswahl gewährleistet wird. Die Antwort ist: Wenig, oder in den Worten von SCHNELL/HILL/ESSER: «[d]ie Verwendung des Begriffs ‚Repräsentativität' ist, legt man wissenschaftliche Kriterien zugrunde, **ungenau und unnötig**»[16].

Externe Validität und Qualität

Die **zentrale Frage**, die es zu beantworten gilt ist, ob von der Stichprobe auf die Grundgesamtheit geschlossen werden kann. Mit anderen Worten, ob die Ergebnisse auf die Grundgesamtheit **verallgemeinerbar** sind (d. h. externe Validität besitzen) oder ob sie nur für die

[13] KRUSKAL/MOSTELLER, a.a.O., S. 13 ff; *dieselben*, Representative Sampling, II: Scientific Literature, Excluding Statistics, ebd., S. 111 ff.; *dieselben*, Representative Sampling, III: The Current Statistical Literature, ebd., S. 245 ff.

[14] Siehe bspw. DIEKMANN, a.a.O., S. 430.

[15] Zit. nach KRUSKAL/MOSTELLER, a.a.O., S. 256.

[16] SCHNELL/HILL/ESSER, a.a.O., S. 298.

Teilmenge der Stichprobe gelten. Die Eignung eines Erhebungsinstruments zeigt sich darin, dass die Charakteristika oder Parameter der Grundgesamtheit (für die relevanten Variablen) möglichst **unverzerrt und effizient geschätzt** werden können.

Keine einzelne Kennzahl kann eine theoretische Analyse zur Qualität der Erhebung ersetzen. Eine umfassende Evaluation erfordert einen **Qualitätsrahmen** (engl. quality framework) und detaillierte Angaben zu Grundgesamtheit, Ziehungsprozess, Art und Umfang der Ausfälle (z. B. infolge von Teilnahmeverweigerung), verwendeten Methoden, usw.[17]

2.3 Qualitätsrahmen

In der Literatur wurden verschiedene Qualitätsrahmen für Erhebungen vorgeschlagen,[18] die sowohl auf nationaler als auch auf supranationaler Ebene umgesetzt wurden. Der **Qualitätsrahmen des European Statistical System** (ESS), dem auch die Amtliche Statistik der Schweiz angeschlossen ist,[19] umfasst den Verhaltenskodex für Europäische Statistiken, den Qualitätssicherungsrahmen des ESS (Quality Assurance Framework) und allgemeine Grundsätze des Qualitätsmanagements.

Der Verhaltenskodex bildet den Eckpfeiler des Qualitätsrahmens. Er enthält 16 Prinzipien für die Entwicklung, Erstellung und Verbreitung von Statistiken.[20] Für jedes Prinzip formuliert der Verhaltenskodex Indikatoren, Standards (Best Practices) und eine Referenz für die Überprüfung der Umsetzung.[21] Einige der Prinzipien sind auf die Amtliche Statistik zugeschnitten (z. B. Vorhandensein eines gesetzlichen Auftrags für die Datenerhebung) und können darum nicht sinnvoll auf nichtstaatliche Einheiten übertragen werden.

Bei der Anwendung des Qualitätsrahmens ESS auf die Rollende Kostenstudie sollen die folgenden Prinzipien des Verhaltenskodex einer Prüfung unterzogen werden: Prinzipien 7 und 8 zum **Einsatz geeigneter statistischer Verfahren und Methoden** und Prinzip 12 **Genauigkeit und Zuverlässigkeit**.

2.4 Kalibrierung

Aus der Literatur zur Nonresponse ist bekannt, dass Item- und Unit-Nonresponse[22] (Teilnahme- oder Antwortverweigerung) nicht zufällig auftreten, sondern selektiv sind.[23] Die

[17] SCHNELL/HILL/ESSER, a.a.O., S. 298.

[18] Siehe bspw. BIEMER/LYBERG, Introduction to Survey Quality, 2003, S. 12 ff; ebenso GROVES ET AL., Survey Methodology, 2009, 2. Aufl., S. 49 ff.

[19] Amtliche Sammlung 2006 5933.

[20] Europäische Union, Verhaltenskodex für Europäische Statistiken, 2018, S. 8 ff.

[21] Eurostat, European Statistical System, Handbook for Quality and Metadata Reports, 2021.

[22] Bei der Nonresponse wird zwischen Item- und Unit-Nonresponse unterschieden. Die Unit-Nonresponse meint die Teilnahmeverweigerung, d. h. den gesamthaften Antwortausfall. Unter Item-Nonresponse wird die Antwortverweigerung zu einzelnen Items oder Fragen verstanden; siehe bspw. SÄRNDAL/LUNDSTRÖM, Estimation in Surveys with Nonresponse, 2005., S. 11 f.

[23] Siehe bspw. BIEMER/LYBERG, Introduction to Survey Quality, a.a.O., S. 63 ff; siehe auch GROVES ET AL., Survey Methodology. 2. Aufl., 2009, S. 183 ff.; ebenso SÄRNDAL/LUNDSTRÖM, a.a.O., S. 11 f.

Nicht-Antwortenden weisen häufig andere Merkmale auf als die Antwortenden. Aus den selektiven Antwortausfällen können verzerrte Schätzungen resultieren.

Die Kalibrierung ist ein statistisches Verfahren, um mögliche Verzerrungen infolge von Coverage Errors[24] und Nonresponse Errors zu verringern (allenfalls zu beheben).[25] Die Kalibrierung wurde ursprünglich für Stichprobenerhebung entwickelt. Sie wird jedoch auch für Nonprobability Samples eingesetzt.[26]

Das Bundesamt für Statistik wendet die Kalibrierung bereits seit Jahren bei einer Vielzahl von Erhebungen an. Sie wird bspw. auch bei der Zensuserhebung «Strukturdaten der Arztpraxen und ambulanten Zentren» (MAS) eingesetzt, um Abweichungen infolge von Coverage und Nonresponse Errors zu reduzieren oder zu korrigieren.[27]

[24] Die Erhebungsgrundgesamtheit umfasst alle Einheiten, die empirisch tatsächlich erhoben werden können. Sie weicht in der Praxis häufig von der (idealtypischen) Ziel-Grundgesamtheit ab. Diskrepanzen zwischen den zwei Formulierungen der Grundgesamtheit werden als Imperfektionen oder Abdeckungsfehler (engl. Coverage Errors) bezeichnet; siehe bspw. SÄRNDAL/LUNDSTRÖM, a.a.O., S. 8.

[25] DEVILLE/SÄRNDAL, Calibration Estimators in Survey Sampling, Journal of the American Statistical Association 87, 1992, S. 376 ff.

[26] Siehe ELLIOTT/VALLIANT, Inference for Nonprobability Samples, in: Statistical Science 32, S. 249 ff; ebenso VALLIANT, Comparing Alternatives for Estimation from Nonprobability Samples, in: Journal of Survey Statistics and Methodology 8, S. 231 ff.

[27] Bundesamt für Statistik, Rapport sur la qualité des données financières. Relevé des données structurelles des cabinets médicaux et des centres ambulatoires MAS, 2022, S. 10.

3 Rollende Kostenstudie

3.1 Charakterisierung der Rollenden Kostenstudie

Das **Forschungsziel** der Rollenden Kostenstudie (RoKo) ist die Schätzung von Kennzahlen oder Parametern (z. B. Mittelwert der Löhne und Sozialabgaben für angestellte Praxisassistentinnen und -assistenten) zur Grundgesamtheit der selbständig praktizierenden Ärztinnen und Ärzte in einem Kanton.[28] Auf Basis der geschätzten Kennzahlen berechnet die Ärztekasse Genossenschaft den Taxpunktwert.

Die RoKo ist eine Datenerhebung mit einem **Querschnittdesign**. Bei einem Querschnittdesign bezieht sich die Datenerhebung auf einen festen Zeitpunkt oder eine kurze Zeitspanne.[29] Die RoKo erfasst Daten für ein bestimmtes Geschäftsjahr. Sie wurde seit 1990 jährlich wiederholt, so dass Querschnitte für eine Vielzahl von Jahren vorliegen.

Die **Erhebungssubjekte** sind die selbständig praktizierenden Ärztinnen und Ärzte in einem Kanton (Spitalambulatorien sind ausgenommen). Die Erhebung der **Individualdaten** (Sachdaten)[30] zu den Ärztinnen und Ärzten wird für jeden Kanton separat durchgeführt.

Die **Erhebungsmethode** ist eine Befragung. Als **Erhebungsinstrument** wird ein Fragebogen mit zwei **Erhebungsmodi** (online und papierbasiert) eingesetzt. Gegenstand der Erhebung sind Angaben zu folgenden (Individual-) **Merkmalen** der selbständig praktizierenden Ärztinnen und Ärzte (Referenz: Geschäftsjahr):[31]

- Aufwände (Material-, Personal-, Raum- und Kapitalaufwand; übriger Aufwand, wie Verwaltungsaufwand, Weiterbildung, Beiträge an Berufsverbände usw.), Abschreibungen;
- Erträge aus ärztlichen und nicht-ärztlichen Leistungen, Medikamenten, Praxislabor, Mittel und Gegenständen/ Material, Miet- und Kapitalerträgen usw.

In Ergänzung zu den finanziellen Daten werden auch Angaben zur Praxis (Grösse, Auslastung, Öffnungszeiten, Angestellte usw.) erhoben.

Mit dem Fragebogen können auch diejenigen Erträge und Aufwände ausgesondert werden, die für **Leistungen nach dem Bundesgesetz über die Krankenversicherung** (KVG) vergütet werden (in Abgrenzung zu den übrigen Leistungen). Für die Berechnung des Taxpunktwertes bilden die ausgesonderten Erträge und Aufwände (siehe Tabelle 1) die zentralen Varia-

[28] Bei den Kennzahlen, Parametern oder Charakteristika handelt es sich um Aggregatmerkmale, die durch Aggregation aus Individualmerkmalen gebildet werden. Als Aggregationsregel wird häufig die Durchschnittsbildung verwendet. Es können auch alternative Masse der zentralen Tendenz, wie Modalwert oder Median benutzt werden; siehe bspw. DIEKMANN, a.a.O., S. 120 f.

[29] Vgl. bspw. DIEKMANN, a.a.O., S. 304 ff.

[30] Es handelt sich um Sachdaten, weil der Personenbezug durch Anonymisierung (irreversibel) aufgehoben wurde.

[31] Ärztekasse, RoKo-Fragebogen 2022, S. 5.

blen. Die Eignung der RoKo zur Berechnung des TPW hängt entscheidend davon ab, ob die Kennzahlen der Grundgesamtheit zuverlässig, fehlerarm und effizient geschätzt werden können.

Tabelle 1 Relevante Variablen zu Aufwand und Ertrag (pro Arzt/ Ärztin und Referenzjahr)

Variable	Bezeichnung	Variable	Bezeichnung
A1T	Total Materialaufwand	A71	Verwaltungsaufwand, Beiträge Berufsverbände, Fortbildung
A2T	Total Personalaufwand	A72	Aufwand aus Praxispartnerschaft
A3T	Total Raumaufwand	A73	Unterhalt und Reparaturen der Praxiseinrichtungen
A4T	Total Kapitalaufwand	A74	Fahrzeugaufwand
A5T	Total Abschreibungen	A75	Praxisversicherungen
B1	Bilanzwerte	E1T	Total Bruttoertrag der Praxistätigkeit

Quelle Ärztekasse, RoKo-Fragebogen/ Übergangsbogen (2017–2022).

Grundgesamtheit

Für die Berechnung des Taxpunktwertes der selbständig praktizierenden Ärztinnen und Ärzten sind nur die KVG-finanzierten Leistungen relevant. Folgerichtig ist die Grundgesamtheit der RoKo in Bezug auf die nach KVG abgerechneten Leistungen definiert.

Die Erbringung von medizinischen Leistungen zulasten der OKP ist für die Ärztinnen und Ärzte an die Erteilung einer kantonalen Berufsausübungsbewilligung gebunden und darum nach Kantonen differenziert zu betrachten. Jede kantonale Ärzteschaft konstituiert eine eigenständige kantonale Grundgesamtheit. Die Zugehörigkeit zu einer kantonalen Grundgesamtheit (für ein bestimmtes Referenzjahr) wird an zwei konstitutiven Merkmalen des TARMED festgemacht. Die Merkmale sind wie folgt definiert.[32]

A1 Die Ärztin/ der Arzt besitzt eine – durch die kantonalen Gesundheitsbehörden erteilte – aktuelle (für das Referenzjahr gültige) Berufsausübungsbewilligung zur Erbringung von OKP-Leistungen.

A2 Die Ärztin/ der Arzt ist:

- reguläres Mitglied der kantonalen Ärztegesellschaft (inkl. Beitritt zum nationalen Rahmenvertrag TARMED und dem kantonalen Tarifvertrag zwischen der Ärztegesellschaft und den Versicherern) oder
- Nicht-Mitglied der kantonalen Ärztegesellschaft und ist dem Rahmen- und Anschlussvertrag des TARMED beigetreten.

Die Bedingungen A1 und A2 bilden gemeinsam eine notwendige und hinreichende Bedingung für die Zugehörigkeit zur (kantonalen) Grundgesamtheit. Nur Ärztinnen und Ärzte, die eines der beiden Teilkriterien unter Lemma A2 erfüllen, können medizinische Leistungen

[32] Siehe dazu SCHOCH ET AL., a.a.O., Kapitel 3.2.

zulasten der OKP nach dem TARMED abrechnen. Folgerichtig ist dieses Merkmal für die Zugehörigkeit zur Grundgesamtheit konstitutiv. Nicht-Mitglieder der kantonalen Ärztegesellschaft (und damit auch nicht der FMH angeschlossen) partizipieren nicht an der Aushandlung der kantonalen Tarifverträge. Sie «schweigen» und schliessen sich dem verhandelten Vertrag an. Vieles spricht dafür, dass es sich um ein qualifiziertes Schweigen oder zustimmendes Schweigen handelt.

Die Grundgesamtheit der RoKo umfasst deshalb die berufstätigen, selbständig praktizierenden Ärztinnen und Ärzte im ambulanten Bereich des Kantons Wallis, die Mitglieder der kantonalen Ärztegesellschaft des Kantons Wallis (SMVS) sind. Die Rahmengrundgesamtheit entspricht dem Mitgliederverzeichnis der SMVS. Von der Rahmengrundgesamtheit wird die Alterskohorte der Über-70-Jährigen ausgeschlossen.[33]

Erhebungsmodus

Die RoKo ist eine Zensuserhebung[34] auf Grundlage des Mitgliederverzeichnisses der SMVS. Die Teilnahme an der RoKo war in den Jahren 2017–2020 freiwillig; entsprechend fiel die Teilnahmerate vergleichsweise gering aus.[35] Im Jahr 2020 betrug die Teilnahmerate 27.3% (d. h. 172 von 629 Ärztinnen und Ärzten haben teilgenommen). Die Teilnahmeraten der anderen Jahre fielen etwa gleich aus. Die Teilmenge der an der RoKo Teilnehmenden besitzt nicht den Charakter einer Zufallsstichprobe (engl. probability sample), sondern wird als Non-probability Sample bezeichnet.

3.2 Nonresponse-Analyse

Durch einen Vergleich der Rahmengrundgesamtheit (Mitgliederverzeichnis, MV) mit den realisierten RoKo-Erhebungsdaten, lassen sich die Teilnahmeraten berechnen und allfällige Diskrepanzen infolge von Nonresponse feststellen.[36] Hierzu werden die vier Variablen: Altersgruppen, Spezialisierung, Geschlecht und Urbanität betrachtet. Die ersten drei Variablen sind selbsterklärend. Die Variable Urbanität (Stadt/Land) unterscheidet ländliche von städtischen Gebieten; für die Zuteilung ist der Ort der Praxis ausschlaggebend. Die Ausprägungen der Variablen sind nachfolgend aufgeführt: `Altersgruppen` (31-40, 41-50, 51-60, 61-70, über 70), `Spezialisierung` (Grund: Grundversorger/innen, Spez: Spezialisten/innen), `Geschlecht` (M: Männer; F: Frauen) und `Urbanität` (Stadt oder Land).

[33] Ein grosser Teil der über-70-jährigen Ärztinnen und Ärzte ist nur noch in kleinem Umfang klinisch tätigt (betreut z. B. nach der regulären Pensionierung sporadisch einen kleinen Patientenstamm). Die gegenüber der OKP abgerechneten Leistungen fallen unterhalb eine Bagatellgrenze von 30 000 CHF.

[34] Bei der Zensuserhebung werden Daten zu allen Erhebungseinheiten (d. h. keine Zufallsauswahl) erfasst; siehe dazu OECD, Glossary of Statistical Terms, 2008.

[35] SCHOCH ET AL., a.a.O., Kapitel 3.2.4.

[36] Die Diskrepanzen könnten auch auf Coverage Errors zurückzuführen sein. Für die anschliessende Kalibrierung ist es unerheblich, ob die festgestellten Abweichungen den Coverage oder Nonresponse Errors zugerechnet werden.

Tabelle 2 Diskrepanzen zwischen den Anteilswerten in der RoKo und dem Mitgliederverzeichnis der SMSV (Jahr 2020)

	Mitgliederverzeichnis SMSV	RoKo	Differenz (Prozentpunkte)
Altersgruppe			
31 - 40 Jahre	10.2%	9.9%	-0.3
41 - 50 Jahre	31.0%	39.5%	8.5
51 - 60 Jahre	31.5%	29.7%	-1.8
61 - 70 Jahre	27.3%	20.9%	-6.4
Spezialisierung			
Grundversorger/in	36.6%	48.3%	11.7
Spezialist/in	63.4%	51.7%	-11.7
Sprache			
Deutsch	22.4%	21.5%	-0.9
Französisch	77.6%	78.5%	0.9
Geschlecht			
Mann	59.3%	53.5%	-5.8
Frau	40.7%	46.5%	5.8
Dimension Stadt-Land			
Ländlich	67.6%	73.3%	5.7
Städtisch	32.4%	26.7%	-5.7

Quelle Mitgliederverzeichnis SMSV und RoKo, Jahr 2020
Anm. Ohne Altersgruppe der Über-70-Jährigen

Die Anteilswerte bzw. Abweichungen zwischen dem MV und den RoKo-Erhebungsdaten (nach Alter, Spezialisierung, Geschlecht und Urbanität) sind in Tabelle 2 ausgewiesen. Lesebeispiel: Die Altersgruppe der 41–50-Jährigen hat nach dem Mitgliederverzeichnis (Rahmengrundgesamtheit) einen Anteil von 31.0 % aller Ärztinnen und Ärzte der Grundgesamtheit. Mit den RoKo-Erhebungsdaten wird der Anteilswert für diese Alterskohorte auf 39.5% geschätzt. Der Anteilswert für die RoKo ist also um 8.5 Prozentpunkte überschätzt. Die Altersgruppe der 41–50-Jährigen ist in der RoKo überrepräsentiert. Aus der Analyse in Tabelle 2 geht überdies hervor, dass die Grundversorger/innen, die Frauen und die ländlichen Gebiete in der RoKo überrepräsentiert sind.

Weitere Anmerkungen

- Eine Über- bzw. Unterrepräsentation impliziert nicht notwendigerweise, dass die Schätzungen für weitere Variablen verzerrt sind.
- Die Tabelle 2 zugrunde liegende Analyse beruht auf einzelnen Variablen (univariate Untersuchung). Sie ist ungeeignet, um Diskrepanzen für Interaktionen von Variablen

zu entdecken (z. B. Kombination von Geschlecht und Spezialisierung). Im nächsten Abschnitt wird eine differenziertere Untersuchung angestrebt.

Anmerkung. Eine weiterführende (technische) Diskussion zu Coverage Errors, Nonresponse Errors, Processing Errors, Measuring Error, etc. im Kontext der RoKo findet sich in SCHOCH ET AL. Dort werden die Daten umfassend untersucht (auch bezüglich Ausreisser und anderer Datenanomalien).[37]

3.3 Kalibrierung

Wie bereits zu Beginn dieses Kapitels erläutert, dient die RoKo als Erhebungsinstrument zur Schätzung von Parametern der Grundgesamtheit, die dann in die Berechnung des TPW einfliessen. Die hierfür relevanten Variablen wurden bereits in Tabelle 1 (oben) aufgeführt. Der Einfluss der Kalibrierung soll nun anhand der relevanten Variablen illustriert werden. Dazu wird der Mittelwert[38] der Variablen:

1. ohne Kalibrierung und
2. mit Kalibrierung (d. h. unter Einbezug der Kalibrierungsgewichte)

berechnet und dann verglichen.

Tabelle 3 Effekt der Kalibrierung auf das geschätzte Mittel der wichtigsten Variablen

Variable	Mittel[1)] (ohne Kalibrierung)	Gewichtetes Mittel[2)] (mit Kalibrierung)	Relative Abweichung	Abweichung in Anz. Std.-Fehler[3)]
A1T	Die Einzelwerte werden nicht veröffentlicht. Sie wurden jedoch den involvierten Behörden zugänglich gemacht.		4.01%	0.45
A2T			1.15%	0.19
A3T			-0.69%	0.11
A4T			-4.77%	0.30
A5T			2.22%	0.21
B1			2.13%	0.15
E1T			1.41%	0.35

Quelle RoKo, Jahr 2020 (ohne und mit kalibrierten Gewichten)

Anm. [1)] 5% gestutztes Mittel; [2)] gewichtetes 5% gestutztes Mittel (Gewichtung gem. Kalibrierung); [3)] Standardfehler des 5% gestutzten Mittels.

[37] SCHOCH ET AL., a.a.O., Kap. 1.5.2, Kap. 3.2, Kap. 3.3 und Anhang B.

[38] Anstelle des arithmetischen Mittels bzw. des gewichteten Mittels, wird das (symmetrisch) 5% gestutzte Mittel (engl. trimmed mean) bzw. das gewichtete 5% gestutzte Mittel berechnet, weil es bezüglich Ausreisser robust ist; siehe SCHOCH ET AL., a.a.O., Anhang B.5.

In Tabelle 3 sind die geschätzten Mittelwerte mit und ohne Kalibrierung festgehalten. Überdies ist die relative Abweichung zwischen den beiden Schätzungen in Prozent ausgewiesen (und die Abweichung in Anzahl Standardfehlern ausgedrückt).

Anmerkung. Die Diskussion der Kalibrierung und alle technischen Details sind im Anhang dokumentiert.

Aus den Schätzungen in Tabelle 3 ist ersichtlich, dass die Abweichungen der geschätzten Mittelwerte mit und ohne Kalibrierung in beide Richtungen auftreten (positive und negative Werte). Die mittlere absolute Abweichung zwischen den Schätzungen beträgt 2.4 Prozentpunkte.[39] Die grösste Diskrepanz (in absoluten Werten) beträgt -4.77% (Variable `A2T`, Total Kapitalaufwand). Das heisst, dass durch die Kalibrierung der geschätzte Mittelwert um 4.77% kleiner ausfällt als ohne Kalibrierung. Für das Total der Bruttoerträge (Variable `E1T`), einer wichtigen Variable für die Berechnung des TPW, beträgt die Diskrepanz zwischen den geschätzten Mittelwerten mit und ohne Kalibrierung nur gerade 1.41%. Sie ist also vernachlässigbar klein. Das gleiche gilt auch für das Total der Aufwände (Summe aus `A1T` bis `A7T`).

Das Ausmass der Diskrepanzen kann auch mithilfe des Standardfehlers des Mittelwertschätzers (zu den Originaldaten) beurteilt werden; siehe letzte Spalte in Tabelle 3. Diese Metrik erlaubt es, die den Daten inhärente Variabilität in der Beurteilung der Diskrepanzen zu berücksichtigen. Im Mittel (über alle relevanten Variablen) haben die Diskrepanzen ein Ausmass von 0.27 Standardfehlern. Ruft man sich in Erinnerung, dass das symmetrische 95%-Vertrauensintervall eine Ausdehnung von knapp +/- 2 Standardfehlern besitzt, so erkennen wir, dass die Diskrepanzen zwischen den ungewichteten und gewichteten (d. h. kalibrierten) Schätzern gering sind (und die Differenzen auf dem 5%-Signifikanzniveau nicht signifikant von null unterschiedlich sind).

Wir halten also fest, dass die Kalibrierung zwar einen Einfluss auf die geschätzten Mittelwerte besitzt, dieser jedoch sehr gering und kaum von Bedeutung ist. Für die RoKo-Daten des Kantons VS erscheint uns – angesichts der geringen Diskrepanzen – eine Kalibrierung im vorliegenden Kontext nicht zwingend.

[39] Die mittlere absoluten Abweichung wurde als arithmetisches Mittel der absoluten Abweichungen berechnet.

4 Schlussfolgerung und Zusammenfassung

Im Kanton Wallis sind die Verhandlungen über den Taxpunktwert (TPW) für die ambulanten ärztlichen Leistungen der frei praktizierenden Ärztinnen und Ärzte gescheitert. Der Staatsrat des Kantons Wallis hat in der Folge den Taxpunktwert für Arztpraxen am 28. November 2018 auf 0.84 CHF festgesetzt. Die Krankenversicherer sowie die Walliser Ärztegesellschaft SMVS haben gegen diesen Entscheid beim Bundesverwaltungsgericht (BVGer) jeweils eigenständig Beschwerde erhoben. Das BVGer hat im Mai 2022 über die Beschwerden entschieden und die Sache an den Staatsrat zurückgewiesen. In seinem Urteil bemängelte das BVGer unter anderem, dass die von der Ärzteschaft für die Berechnung des TPW herangezogene Rollende Kostenstudie (RoKo) nicht «repräsentativ» und somit für die Berechnung des TPW ungeeignet sei.

Das BVGer bezieht sich auf ein **metaphorisches Verständnis** des Begriffs «Repräsentativität», wonach eine Erhebung dann «repräsentativ» sei, wenn die «Stichprobe ein verkleinertes, aber sonst wirklichkeitsgetreues Abbild der Grundgesamtheit» darstelle. Das Gericht lässt allerdings offen, nach welchen Kriterien zu beurteilen ist, ob eine wirklichkeitsgetreue Abbildung vorliegt.

Entgegen einer weit verbreiteten Meinung sind «Repräsentativität» und «repräsentative Stichprobe» **keine Begriffe der Stichprobentheorie** (oder der Statistik im Allgemeinen). Der Begriff «Repräsentativität» ist keine quantifizierbare Eigenschaft einer Stichprobe und wird in der wissenschaftlichen Literatur (mit wenigen Ausnahmen) nicht verwendet. In der Alltags- und Verwaltungssprache zeichnet sich der Wortgebrauch von «Repräsentativität» durch ein hohes Mass an **Ambiguität** aus. Der mediale Diskurs über «repräsentative Stichproben» wird vor allem von Markt- und Meinungsforschungsinstitute dominiert, die ihren Produkten die Aura der Wissenschaftlichkeit verleihen wollen. Der Begriff ist allerdings kaum mehr als «schmückendes und vorwiegend inhaltsleeres Attribut».

Die **zentrale Frage**, die es zu beantworten gilt ist, ob **von der Stichprobe auf die Grundgesamtheit geschlossen** werden kann. Mit anderen Worten, ob die Ergebnisse auf die Grundgesamtheit verallgemeinerbar sind (d. h. **externe Validität** besitzen) oder ob sie nur für die Teilmenge der Stichprobe gelten. Die Eignung eines Erhebungsinstruments zeigt sich darin, ob die Charakteristika oder Parameter der Grundgesamtheit (für die relevanten Variablen) möglichst **unverzerrt und effizient geschätzt** werden können. Eine umfassende Evaluation erfordert einen Qualitätsrahmen (engl. quality framework) und detaillierte Angaben zu Grundgesamtheit, Ziehungsprozess, Art und Umfang der Ausfälle, verwendeten Methoden, usw.

Der **Qualitätsrahmen des European Statistical System**, dem auch die Amtliche Statistik der Schweiz angeschlossen ist, wird als geeignet erachtet, die Qualität der RoKo (auch hinsichtlich externer Validität) in einer empirischen Analyse zu beurteilen.

Die **empirische Analyse** zur RoKo hat gezeigt, dass es mitunter grosse Diskrepanzen bei den Anteilswerten zwischen den RoKo-Erhebungsdaten und dem kantonalen Mitgliederverzeich-

nis (nach Alter, Spezialisierung, Geschlecht und Urbanität) gibt. Aus dem Vergleich geht hervor, dass die Altersgruppe der 41-50-Jährigen, die Grundversorger/innen, die Frauen und die ländlichen Gebiete in der RoKo überrepräsentiert sind. Die Diskrepanzen sind mehrheitlich auf Verzerrungen infolge von Nonresponse zurückzuführen.

Um die Diskrepanzen genauer abzuschätzen und deren Effekt auf die Schätzung der relevanten Variablen zu beurteilen, wurde eine **Kalibrierung** für die RoKo durchgeführt (bezüglich der Variablen: Spezialisierung, Alter, Geschlecht, Stadt/Land und Sprache). Die Kalibrierung ist ein Verfahren, um mögliche Verzerrungen infolge von Coverage und Nonresponse Error zu verringern (allenfalls zu beheben). Nach der Kalibrierung wurden die geschätzten Mittelwerte der relevanten Variablen ungewichtet (Originaldaten) und mit den kalibrierten Gewichten berechnet, um sie anschliessend miteinander zu vergleichen. Der Vergleich zeigte auf, dass die mittlere absolute Abweichung zwischen den ungewichteten und gewichteten Mittelwerten 2.4 Prozentpunkte beträgt. Das Ausmass der Diskrepanzen kann auch mithilfe des Standardfehlers des Mittelwertschätzers beurteilt werden. Diese Metrik erlaubt es, die den Daten inhärente Variabilität in der Beurteilung der Diskrepanzen zu berücksichtigen. Im Mittel (über alle relevanten Variablen) haben die Diskrepanzen ein Ausmass von 0.27 Standardfehlern.

Unabhängig davon, auf welcher Grundlage die Abweichungen gemessen werden, kann festgestellt werden, dass die **Unterschiede zwischen den ungewichteten und den (nach der Kalibrierung) gewichteten Mittelwerten** (für die relevanten Variablen) **sehr gering** sind. Es ist daher festzuhalten, dass die Kalibrierung zwar einen Einfluss auf die geschätzten Mittelwerte hat, dieser jedoch sehr gering und kaum bedeutsam ist. Für die RoKo-Daten des Kantons VS erscheint eine Kalibrierung im vorliegenden Kontext nicht zwingend. Die Schätzung von Charakteristika und Parameter der Grundgesamtheit (für die relevanten Variablen) ist, bezüglich der Kalibrierungsvariablen (Spezialisierung, Alter, Geschlecht, Stadt/Land und Sprache) mit grosser Wahrscheinlichkeit nicht verzerrt. Diese empirisch fundierte Feststellung steht in klarem Widerspruch zum Einwand des Staatsrates und des BVGer, die RoKo sei nicht «repräsentativ».

Literaturverzeichnis

AGRESTI, A. (2002). Categorical Data Analysis. Hoboken (NJ): John Wiley & Sons, 2. Aufl.

BENESCH, T. (2013). Schlüsselkonzepte zur Statistik: Die wichtigsten Methoden, Verteilungen, Tests anschaulich erklärt. Heidelberg: Spektrum Akademischer Verlag.

BFS – Bundesamt für Statistik (2022). Rapport sur la qualité des données financières. Relevé des données structurelles des cabinets médicaux et des centres ambulatoires MAS, Office fédéral de la statistique, Neuchâtel (Numéro OFS be-f-14.04.05-03).

BIEMER, P. P. und LYBERG, L. E. (2003). Introduction to Survey Quality. Hoboken (NJ): John Wiley & Sons.

DEVILLE, J.-C. und SÄRNDAL, C.-E. (1992). Calibration Estimators in Survey Sampling, Journal of the American Statistical Association 87, S 376–382. Doi: https://doi.org/10.1080/01621459.1992.10475217

DEVILLE, J.-C., SÄRNDAL, C.-E. und SAUTORY, O. (1993). Generalized Raking Procedures in Survey Sampling. Journal of the American Statistical Association 88, S. 1013–1020. Doi: 10.1080/01621459.1993.10476369

DIEKMAN, A. (2021). Empirische Sozialforschung. Grundlagen, Methoden, Anwendungen, 14. Aufl., Reinbek bei Hamburg: Rowohlt.

ELLIOTT, M. R. und VALLIANT, R. (2017). Inference for Nonprobability Samples. Statistical Science 32, S. 249–264. Doi: https://doi.org/10.1214/16-STS598

Eurostat (2021). European Statistical System. Handbook for Quality and Metadata Reports, 2021 re-edition, Luxemburg. Doi: https://doi.org/10.2785/666412

GROVES, R. M. und LYBERG, L. (2010). Total Survey Error: Past, Present, and Future. Public Opinion Quarterly 74, S. 849–879. Doi: https://doi.org/10.1093/poq/nfq065

GROVES, R. M., FOWLER, F. J., COUPER, M. P., LEPKOWSKI, J. M., SINGER, E. und TOURANGEAU, R. (2009). Survey Methodology. 2. Aufl. Hoboken (NJ): John Wiley & Sons.

KRUSKAL, W. H. und MOSTELLER, F. (1979a). Representative Sampling, I: Non-Scientific Literature. International Statistical Review 47, S. 13–24. Doi: https://doi.org/10.2307/1403202

KRUSKAL, W. H. und MOSTELLER, F. (1979b). Representative Sampling, II: Scientific Literature, Excluding Statistics. International Statistical Review 47, S. 111–127. Doi: https://doi.org/10.2307/1402564

KRUSKAL, W. H. und MOSTELLER, F. (1979c). Representative Sampling, III: The Current Statistical Literature. International Statistical Review 47, S. 245–265. Doi: https://doi.org/10.2307/1402647

LITTLE, R. J. und VARTIVARIAN, S. (2005). Does Weighting for Nonresponse Increase the Variance of Survey Means? Survey Methodology 31, S. 161–168.

LUMLEY, T. (2010). Complex Surveys: A Guide to Analysis Using R. Hoboken (NJ): John Wiley & Sons.

LUMLEY, T. (2023). survey: Analysis of complex survey samples. R package version 4.2. URL: https://CRAN.R-project.org/package=survey

OECD – Organisation for Economic Co-operation and Development (2008). Glossary of Statistical Terms. Paris: OECD. Doi: 10.1787/9789264055087-en

PÖRKSEN, U. (2011). Plastikwörter: Die Sprache einer internationalen Diktatur. Stuttgart: Klett-Cotta, 7. Aufl.

R Development Core Team (2023). R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria. URL: http://www.R-project.org

SÄRNDAL, C.-E. und LUNDSTRÖM, S. (2005). Estimation in Surveys with Nonresponse. Hoboken (NJ): John Wiley & Sons.

SCHNELL, R., HILL, P. B. und ESSER, E. (2018). Methoden der empirischen Sozialforschung. Berlin: De Gruyter Oldenbourg, 11., überarb. Aufl.

SCHOCH, T., HULLIGER, B., SPASOVA, T. und THEES, O. (2023). Evaluation der Datengrundlagen und der Datenaufbereitung für das Kostenmodell KOREG, Olten. Studie im Auftrag der ats-tms AG.

TILLÉ, Y. (2019). Théorie des Sondages: Échantillonnage et Estimation en Populations Finies. 2. Aufl., Paris: Dunod.

VALLIANT, R. (2020). Comparing Alternatives for Estimation from Nonprobability Samples. Journal of Survey Statistics and Methodology 8, S. 231–263. Doi: https://doi.org/10.1093/jssam/smz003

Quellenverzeichnis

Urteil des Bundesverwaltungsgerichts C-7338/2018 vom 20. Mai 2022 (Décision du Conseil d'Etat du canton du Valais du 28 novembre 2018 fixant la valeur du point des prestations ambulatoires TARMED extrahospitalières fournies dès le 1er janvier 2017 à la charge des assureurs recourants.)

Sachdaten der Rollenden Kostenstudie für den Kanton Wallis (Ärztekasse Genossenschaft, Jahre 2017-2020)

Mitgliederverzeichnis SMVS (Jahre 2017-2020)

Anhang

Die Kalibrierung der RoKo-Daten auf das Mitgliederverzeichnis der SMVS (siehe Tabelle 4) beruht auf der raking-Distanzfunktion[40]. Die Berechnung der kalibrierten Gewichte erfolgt nach dem Algorithmus von DEVILLE/SÄRNDAL/SAUTORY[41], der als Funktion *grake()* im R-package survey implementiert ist.[42] Als Ausgangsgewichte wurden Gewichte der Grösse 1 verwendet.

Tabelle 4 Mitgliederverzeichnis der SMSV (Jahr 2020)

	Altersgruppe		31-40		41-50		51-60		61-70	
	Geschlecht		M	W	M	W	M	W	M	W
Spezialisierung	Stadt/ Land	Sprache								
Grundversorg.	Land	D	2	-	5	9	15	5	13	5
		F	10	12	13	17	21	22	30	8
	Stadt	D	-	-	-	-	-	1	-	-
		F	1	6	8	9	3	6	8	1
Spezialist/in	Land	D	3	3	17	10	15	10	18	2
		F	4	13	26	27	22	20	38	10
	Stadt	D	-	-	1	2	4	-	1	-
		F	3	7	26	25	36	18	30	8

Quelle RoKo/ SMSV, 2020
Anm. Ohne Altersgruppe der Über-70-Jährigen; leere Zellen sind mit einem Strich (-) gekennzeichnet.

Kalibrierungsbedingungen

Für die Kalibrierung stehen folgende kategorischen Variablen zur Verfügung:

- Altersgruppen (31-40, 41-50, 51-60 oder 61-70)
- Spezialisierung (Grund: Grundversorger/innen, Spez: Spezialisten/innen)
- Geschlecht (M: Männer, F: Frauen)
- Stadt/Land (Stadt oder Land)
- Sprache (D: Deutsch, F: Französisch)

Die Ausprägungen einer Variable werden im folgenden in der Form `[Variable]_[Ausprägung]` geschrieben; z. B. der Ausdruck `Spezialisierung_Spez` meint die Ausprägung «Spe-

[40] DEVILLE/SÄRNDAL, 1992, a.a.O., Fall 2 in Tabelle 1.
[41] DEVILLE/SÄRNDAL/SAUTORY, Generalized Raking Procedures in Survey Sampling, in: Journal of the American Statistical Association 88/1993, S. 1018.
[42] LUMLEY, Complex Surveys: A Guide to Analysis Using R, 2010; *derselbe*, survey: Analysis of Complex Survey Samples, 2023, S. 139 ff.

zialist/in» der Variable Spezialisierung. Interaktionsterme zwischen zwei Ausprägungen sind durch einen Doppelpunkt (:) gekennzeichnet, z. B. `Sprache_F : Geschlecht_M` bezeichnet die Interaktion 2. Ordnung der Ausprägungen Französisch (Sprache) und Mann (Geschlecht). Es können auch Interaktionen höherer Ordnung auftreten. Insgesamt resultieren 64 Interaktionsterme (Kombinationsmöglichkeiten).

Mithilfe von log-linearen Modellen[43] wurden die Daten des Mitgliederverzeichnisses SMVS modelliert, um die Anzahl der Interaktionsterme zu reduzieren, die dann anschliessend als Bedingungen für die Kalibrierung verwendet werden sollen. Damit konnten die folgenden 28 Interaktionsterme 2. Ordnung[44] identifiziert werden:

```
Spezialisierung_Grund : Altersgruppe_31-40
Spezialisierung_Spez : Altersgruppe_31-40
Spezialisierung_Grund : Altersgruppe_41-50
Spezialisierung_Spez : Altersgruppe_41-50
Spezialisierung_Grund : Altersgruppe_51-60
Spezialisierung_Spez : Altersgruppe_51-60
Spezialisierung_Grund : Altersgruppe_61-70
Spezialisierung_Spez : Altersgruppe_61-70
Spezialisierung_Grund : StadtLand_Land
Spezialisierung_Spez : StadtLand_Land
Spezialisierung_Grund : StadtLand_Stadt
Spezialisierung_Spez : StadtLand_Stadt
Sprache_D : Geschlecht_M
Sprache_F : Geschlecht_M
Sprache_D : Geschlecht_W
Sprache_F : Geschlecht_W
Geschlecht_M : Altersgruppe_31-40
Geschlecht_W : Altersgruppe_31-40
Geschlecht_M : Altersgruppe_41-50
Geschlecht_W : Altersgruppe_41-50
Geschlecht_M : Altersgruppe_51-60
Geschlecht_W : Altersgruppe_51-60
Geschlecht_M : Altersgruppe_61-70
Geschlecht_W : Altersgruppe_61-70
StadtLand_Land : Sprache_D
StadtLand_Stadt : Sprache_D
StadtLand_Land : Sprache_F
StadtLand_Stadt : Sprache_F
```

Die oben aufgeführten 28 Interaktionsterme wurden anschliessend für die Daten des Mitgliederverzeichnisses SMVS als Dummy-Variable (d. h. 0-1-wertige Indikatorvariable) formu-

[43] Siehe bspw. AGRESTI, Categorical Data Analysis, 2. Aufl., 2002, S. 314 ff.

[44] Der Erklärungsgehalt von Interaktionstermen höherer Ordnung war in den log-linearen Modelle marginal.

liert, um zu prüfen, ob bzw. welche Interaktionsterme kollinear sind. Nach Ausschluss der kollinearen Dummy-Variablen resultierten die folgenden 17 Interaktionsterme:

```
Spezialisierung_Grund : Altersgruppe_31-40
Spezialisierung_Spez : Altersgruppe_31-40
Spezialisierung_Grund : Altersgruppe_41-50
Spezialisierung_Spez : Altersgruppe_41-50
Spezialisierung_Grund : Altersgruppe_51-60
Spezialisierung_Spez : Altersgruppe_51-60
Spezialisierung_Grund : Altersgruppe_61-70
Spezialisierung_Spez : Altersgruppe_61-70
Spezialisierung_Grund : StadtLand_Land
Spezialisierung_Spez : StadtLand_Land
Sprache_D : Geschlecht_M
Sprache_F : Geschlecht_M
Sprache_D : Geschlecht_W
Geschlecht_M : Altersgruppe_31-40
Geschlecht_M : Altersgruppe_41-50
Geschlecht_M : Altersgruppe_51-60
StadtLand_Land : Sprache_D
```

Die 17 resultierenden Interaktionsterme wurden anschliessend zu den Bedingungen (engl. calibration constraints) für die Kalibrierung erklärt.

Kennzahlen zur Kalibrierung

Das der Kalibrierung unterliegende Optimierungsproblem wird durch `grake()` nach 11 Iterationsschritten gelöst (Termination rule: Euklidische Norm der Gewichtsanpassung ist kleiner als 10^{-7}). Das kleinste kalibrierte Gewicht hat einen Wert von 1.92; das grösste Gewicht beträgt 9.60 (siehe Tabelle 5). Verhältnis zwischen dem grössten und dem kleinsten kalibrierten Gewicht beträgt knapp 5. Es treten weder negative Gewichte noch ausnehmend grosse Gewichte auf.

Tabelle 5 Five-number summary (und arithmetisches Mittel) der kalibrierten Gewichte

Minimum	25%-Perzentil	Median	Mittelwert	75%-Perzentil	Maximum
1.92	2.72	3.14	3.66	4.04	9.60

Quelle RoKo Kanton VS, 2020 (ohne Altersgruppe der Über-70-Jährigen).

Ziel der Kalibrierung ist es, die (potenzielle) Nonresponse-Verzerrung zu reduzieren. Die für die Kalibrierung verwendeten Hilfsinformationen (Kalibrierungsvariablen) sollten zwei Eigenschaften aufweisen, um die Nonresponse-Verzerrung zu reduzieren[45]. Die Hilfsinformation muss

[45] LITTLE/VARTIVARIAN, Does Weighting for Nonresponse Increase the Variance of Survey Means?, in: Survey Methodology 31/2005, S. 4 ff.

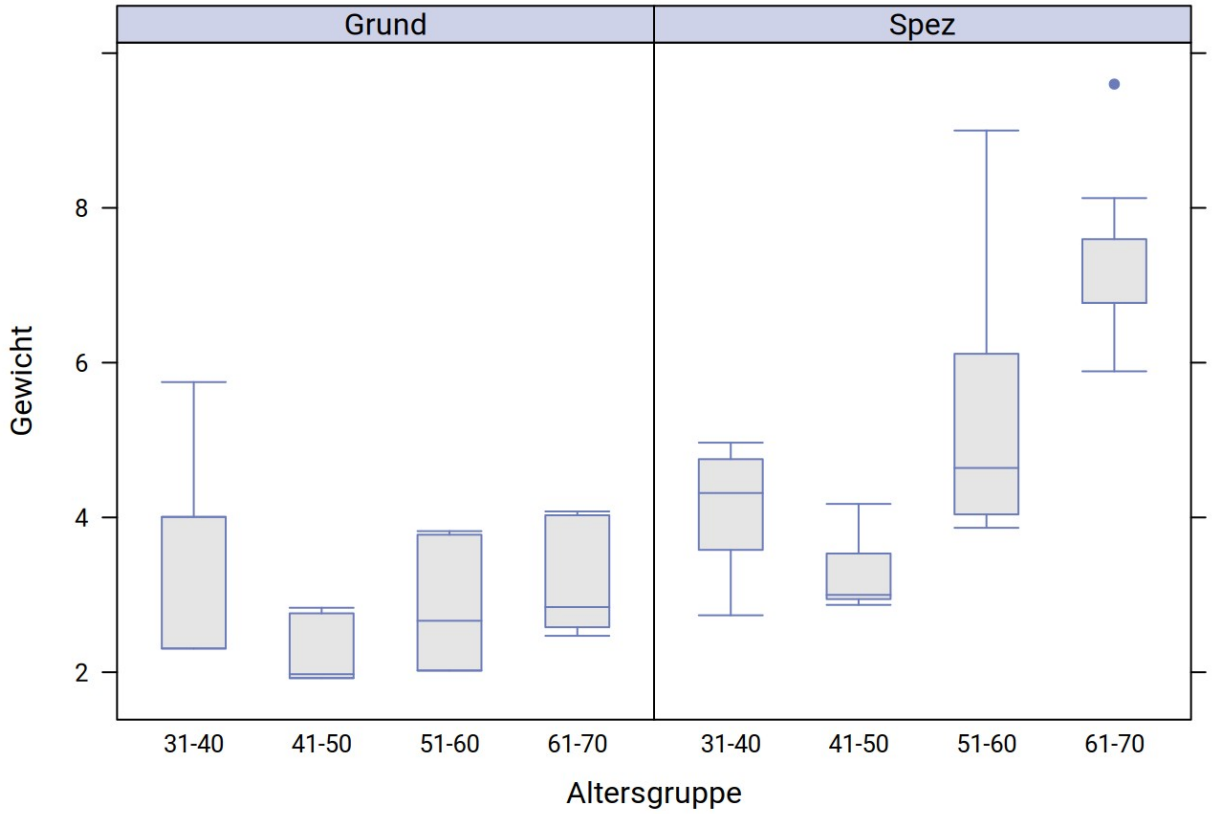
1. mit der Antwortwahrscheinlichkeit korreliert sein und
2. für die interessierende(n) Variable(n) prädiktiv sein.

Wenn letzteres der Fall ist, führt die Kalibrierung/ Gewichtung (tendenziell) zu einer geringeren Nonresponse-Verzerrung und einer geringeren Stichprobenvarianz. Falls 2) nicht zutrifft, so führt eine Kalibrierung zu einer Zunahme der Stichprobenvarianz. Aus diesen Überlegungen ist es also nicht ratsam, alle verfügbaren Hilfsinformationen in der Kalibrierung zu verwenden. Wir halten die Auswahl der Hilfsinformationen/ Kalibrierungsvariablen für angemessen.

Analyse der Gewichte nach den Variablen Spezialisierung, Stadt/Land und Geschlecht

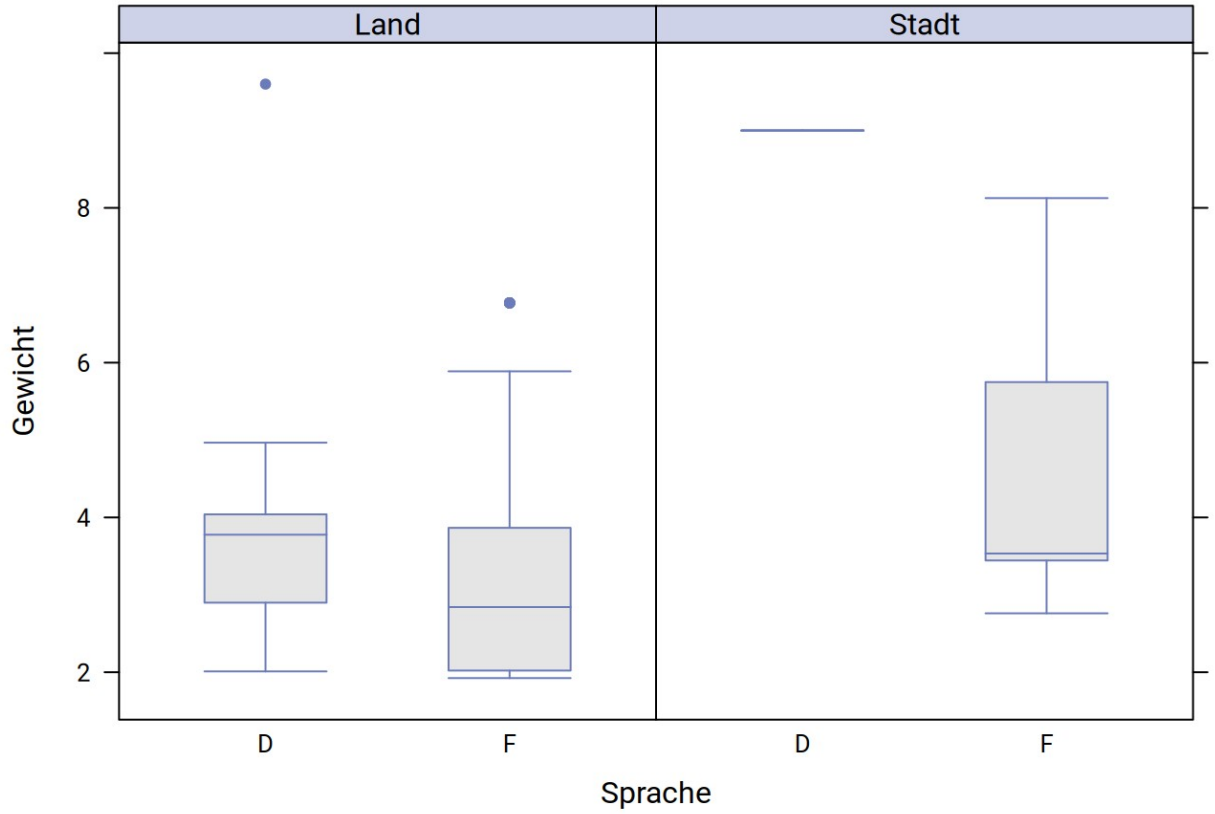
Die Abbildungen 1–3 zeigen die Verteilung der kalibrierten Gewichte nach Altersgruppe und Spezialisierung, Stadt/Land und Sprache, sowie Geschlecht und Stadt/Land.

Abbildung 1 Boxplot der kalibrierten Gewichte nach Altersgruppe und Spezialisierung



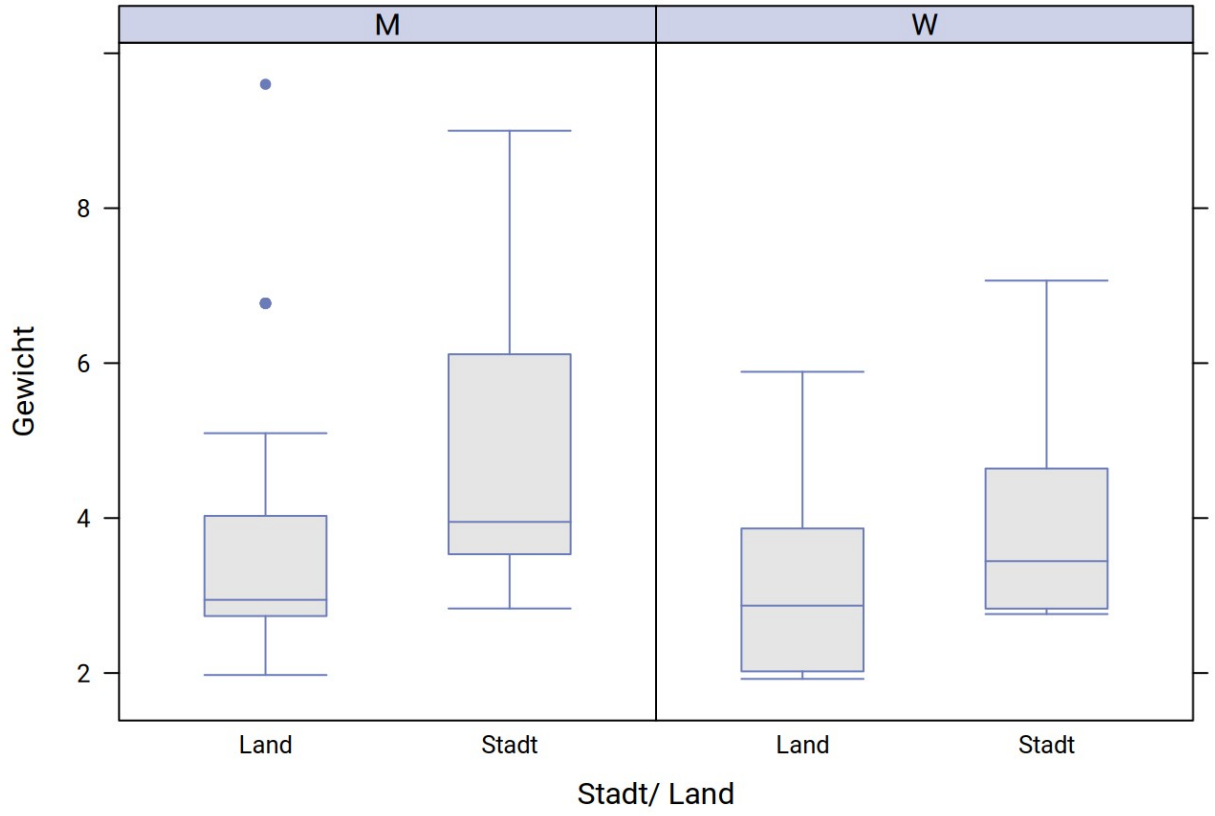
Quelle RoKo, 2020.

Abbildung 2 Boxplot der kalibrierten Gewichte nach Sprache und Stadt/Land



Quelle RoKo, 2020.

Abbildung 3 Boxplot der kalibrierten Gewichte nach Stadt/Land und Geschlecht



Quelle RoKo, 2020.