

Fachhochschule Nordwestschweiz, Riggenbachstr. 16, 4600 Olten

Geschäftsstelle PROphysio  
Frau Nina Zosso, Präsidentin ad interim  
Impact2go  
Mattenhofstrasse 5  
CH-3007 Bern

Olten, 26.10.2026

### **Erwiderung auf das Gutachten von Polynomics zur LeDa-Studie (offener Brief)**

Sehr geehrte Frau Zosso

Das von PROphysio in Auftrag gegebene Gutachten von Polynomics<sup>1</sup> übt in verschiedenen Punkten Kritik an unserer LeDa-Studie<sup>2</sup>. Im Namen der Arbeitsgemeinschaft Ecoplan AG/FHNW nehmen wir hierzu mit der vorliegenden Erwiderung aus methodischer und empirischer Sicht Stellung. Nachfolgend gehen wir auf die in Kapitel 3 des Gutachtens geäusserten Kritikpunkte ein.

1. Unterschiede der Zeitdauern nach Praxisgrösse
2. Rahmenpopulation und Schichtung nach Praxisgrösse
3. Vermeintliche Mängel bei der statistischen Auswertung

Wir danken Ihnen für die Berücksichtigung der gemachten Ausführungen und stehen Ihnen für weitere Auskünfte gerne zur Verfügung.

Mit freundlichen Grüssen

Prof. Dr. Tobias Schoch  
Stv. Institutsleiter ICC FHNW

Eliane Kraft  
Geschäftsführerin Ecoplan AG

---

<sup>1</sup> Polynomics, Gesundheitsökonomische Prüfung der neuen KVG-Tarifstruktur für die Physiotherapie Studie zu Kosten und Leistungen in der Physiotherapie. Olten, 2026 – Gutachten im Auftrag von PROphysio (Fassung vom 10. Sept. 2026), siehe: [https://www.polynomics.ch/admin/data/files/publication/document/448/polynomics\\_prufung\\_kvg-tarif\\_physio\\_gutachten\\_2026-09-10\\_korr.pdf](https://www.polynomics.ch/admin/data/files/publication/document/448/polynomics_prufung_kvg-tarif_physio_gutachten_2026-09-10_korr.pdf), ebenso <https://prophysioverband.ch/unabhaengiges-forschungsinstitut-bestaetigt-der-neue-tarif-weist-maengel-auf>.

<sup>2</sup> Ecoplan/FHNW, Studie zu den Leistungen in der Physiotherapie. Bern/Olten, 2025 – Studie im Auftrag von Physioswiss (Fassung vom 22. Okt. 2025).

## Kurzfassung/Zusammenfassung

Polynomics regt eine zusätzliche Auswertung der Behandlungsdauern nach Praxisgrösse an. Dieser Anregung wird hier entsprochen. Die Behandlungsdauer nimmt mit zunehmender Praxisgrösse zwar leicht ab. Ein designbasierter globaler Wald-Test unter Berücksichtigung des Stichprobenplans zeigt jedoch keine statistisch signifikanten Unterschiede zwischen kleinen, mittleren und grossen Praxen. Die Daten liefern damit keine hinreichende statistische Evidenz für Unterschiede zwischen den Praxisgrössen.

### *Rahmenpopulation und Schichtung nach Praxisgrösse*

Die Praxisgrösse wurde in der LeDa-Studie anhand der im Zahlstellenregister (ZSR) verfügbaren K-Nummern definiert. Da K-Nummern die tatsächliche Zahl der beschäftigten Physiotherapeut/innen nicht durchgehend vollständig abbilden, ist eine Fehlklassifikation einzelner Praxen möglich. Die K-Nummern waren jedoch das einzige im ZSR verfügbare Merkmal, das Rückschlüsse auf die Praxisgrösse erlaubte; bei gegebenem Stichprobenrahmen bestand die unmittelbar verfügbare Alternative somit in der Nutzung dieser approximativen Information oder im Verzicht auf die Praxisgrösse als Schichtungsmerkmal.

Soweit Polynomics eine mögliche Verzerrung aus der fehlerhaften Zuordnung von Praxen zu Grössenklassen innerhalb des Stichprobenrahmens herleitet, ist dieser Wirkungsmechanismus theoretisch nicht begründet. Die Stichprobenliteratur zeigt, dass eine solche Fehlklassifikation bei korrekter Berücksichtigung des tatsächlich realisierten Stichprobenplans nicht bereits als solche zu einer systematischen Verzerrung der designbasierten Schätzung führt. Davon zu unterscheiden wäre eine Unterabdeckung der Zielpopulation durch das ZSR, die grundsätzlich zu Verzerrungen führen könnte; hierfür liegen aus unserer Sicht jedoch keine Anhaltspunkte vor.

Der von Polynomics aus einer fehlerhaften Schichtzuweisung abgeleitete Verzerrungsmechanismus lässt sich somit nicht herleiten.

### *Sensitivitätsanalyse*

Zur quantitativen Einordnung der von Polynomics thematisierten möglichen Abweichungen wurden ergänzend Sensitivitätsanalysen durchgeführt. Diese quantifizieren nicht eine durch die approximative Bestimmung der Praxisgrösse mittels K-Nummern verursachte Verzerrung; wie oben erläutert, folgt aus einer fehlerhaften Schichtzuweisung allein kein entsprechender systematischer Bias. Untersucht wird vielmehr kontrafaktisch, wie stark sich der geschätzte Gesamtmittelwert verändern würde, wenn der Anteil grosser Praxen tatsächlich höher wäre. Der von Polynomics aus dem GLN-Register abgeleitete Anteil grosser Praxisstandorte von 9% wird dabei ausdrücklich nur als alternatives Szenario verwendet. Unter dieser Annahme sinkt die geschätzte mittlere Behandlungsdauer gegenüber dem LeDa-Wert um lediglich rund 5.8 Sekunden (von 34.24 auf 34.15 Minuten) bzw. weniger als 0.3%. Auch weitergehende, bewusst ungünstige Stressszenarien führen nur zu Reduktionen von rund 10 bis 22 Sekunden. Die quantitative Bedeutung der möglichen Abweichung ist unter den betrachteten Szenarien daher gering.

### *Vermeintliche Mängel bei der statistischen Auswertung*

Die von Polynomics beanstandete Diskrepanz bei der statistischen Signifikanz beruht darauf, dass der Begutachtung nicht die finale Fassung des LeDa-Berichts zugrunde lag. In der finalen Version wird der Unterschied zwischen selbständig und angestellt tätigen Physiotherapeut/innen auf Grundlage eines designbasierten Tests als statistisch signifikant ausgewiesen.

Ergänzend ist festzuhalten, dass die von Polynomics vorgenommene Nachrechnung keine korrekte Rekonstruktion der im LeDa-Bericht verwendeten designbasierten Inferenz darstellt. Polynomics kombiniert die ausgewiesenen designbasierten Standardfehler mit den nominellen Gruppenumfängen im Rahmen einer klassischen Welch-Satterthwaite-Approximation. Diese Vorgehensweise berücksichtigt die Freiheitsgrade des komplexen Stichprobendesigns nicht. Soweit sich aus dieser Nachrechnung Abweichungen gegenüber den im LeDa-Bericht ausgewiesenen inferenzstatistischen Ergebnissen ergeben, können diese daher nicht als Hinweis auf eine fehlerhafte oder widersprüchliche Auswertung der LeDa-Studie interpretiert werden. Sie sind vielmehr auf die von Polynomics gewählte, für das dokumentierte Stichprobendesign nicht adäquate Rekonstruktionsmethode zurückzuführen.

## 1. Unterschiede der Zeitdauern nach Praxisgrösse

Polynomics regt an, die Behandlungsdauern nach Praxisgrösse auszuwerten, um die Hypothese kürzerer Behandlungszeiten in grösseren Praxen direkt zu prüfen.<sup>3</sup> Dieser Forderung wird nachfolgend entsprochen.

Die geschätzte mittlere Behandlungsdauer nimmt mit zunehmender Praxisgrösse ab (Tabelle 1; LeDa-Erhebung, Selbstbeobachtung). Sie beträgt 34.41 Minuten in kleinen, 33.49 Minuten in mittleren und 32.95 Minuten in grossen Praxen. Der Unterschied zwischen kleinen und grossen Praxen beträgt damit rund 1.46 Minuten bzw. 4.2% bezogen auf den Mittelwert der kleinen Praxen.

| Praxisgrösse | Mittelwert | KI (unten) | KI (oben) |
|--------------|------------|------------|-----------|
| klein        | 34.41      | 33.34      | 35.48     |
| mittel       | 33.49      | 32.24      | 34.74     |
| gross        | 32.95      | 31.27      | 34.63     |

**Tabelle 1:** Geschätzte mittlere Behandlungsdauer (Min.) nach Praxisgrösse und 95%-Konfidenzintervalle (KI); Daten: LeDa-Studie, Selbstbeobachtung

Um die Unterschiede zwischen den drei Praxisgrössenklassen inferenzstatistisch zu untersuchen, wurde ergänzend ein designbasierter Wald-Test durchgeführt.<sup>4</sup> Geprüft wurde die Nullhypothese, dass die mittlere Behandlungsdauer in kleinen, mittleren und grossen Praxen übereinstimmt. Der Wald-Test ergibt  $F(2, 294) = 1.23$  bei  $p = 0.294$ . Die Nullhypothese kann somit auf den üblichen Signifikanzniveaus nicht verworfen werden. Die vorstehenden Auswertungen nach Praxisgrösse wurden im LeDa-Bericht nicht gesondert ausgewiesen. Die Punktschätzer weisen zwar in die von Polynomics vermutete Richtung, der globale designbasierte Wald-Test liefert jedoch keine hinreichende statistische Evidenz für Unterschiede zwischen den Praxisgrössenklassen.<sup>5</sup>

## 2. Rahmenpopulation und Schichtung nach Praxisgrösse

Die Rahmenpopulation der LeDa-Studien basiert auf dem Zahlstellenregister (ZSR) der SASIS AG (heute santéservices AG). Die Praxisgrösse wurde anhand der im ZSR verfügbaren Angaben zu den K-Nummern bestimmt. Wie Polynomics zutreffend festhält,<sup>6</sup> besteht keine gesetzliche oder tarifvertragliche Pflicht, für angestellte Physiotherapeut/innen K-Nummern zu beziehen und zu führen.

Die Vergabe und Führung von K-Nummern erfolgt in der Praxis nicht einheitlich. Es ist daher davon auszugehen, dass die Anzahl der K-Nummern die tatsächliche Zahl der in einer Praxis beschäftigten Physiotherapeut/innen nicht durchgehend vollständig und zutreffend abbildet. Das Merkmal kann folglich lediglich als *approximative* Information zur Praxisgrösse herangezogen werden. Die Praxisgrösse wurde in der LeDa-Studie als *Schichtungsmerkmal* für die Stichprobenziehung verwendet. Für die Beurteilung dieser Entscheidung ist wesentlich, dass die Anzahl der erteilten K-Nummern das *einzige* im ZSR verfügbare Merkmal war, das überhaupt Rückschlüsse auf die Praxisgrösse zulies.<sup>7</sup>

<sup>3</sup> Polynomics, S. 15–16 und S. 18.

<sup>4</sup> Hierzu wurde die Funktionalität des R-Pakets "survey" verwendet, wie in der LeDa-Studie dokumentiert; vgl. Eco-plan/FHNW, S. 44. Eine ausführliche Darstellung der mit dem Paket "survey" umgesetzten Methoden findet sich bei Lumley, T., *Complex Surveys: A Guide to Analysis Using R*. Hoboken (NJ): John Wiley & Sons, 2010.

<sup>5</sup> Ergänzend zum globalen designbasierten Wald-Test wurde der Kontrast zwischen kleinen und grossen Praxen untersucht. Der geschätzte Unterschied beträgt 1.46 Minuten zugunsten kürzerer Behandlungsdauern in grossen Praxen (95%-KI: [−0.53; 3.45] Minuten). Der entsprechende Kontrast ist statistisch nicht signifikant ( $p = 0.152$ ).

<sup>6</sup> Polynomics, S. 13–14.

<sup>7</sup> Die in der ZSR-Datenlieferung enthaltenen Merkmale bzw. Variablen sind in Eco-plan/FHNW, S. 38 dokumentiert. Hervorzuheben ist, dass SASIS AG die Identifikatoren GLN und UID Eco-plan/FHNW trotz entsprechender Anfrage nicht zur Verfügung stellte. Dadurch war insbesondere eine eindeutige Verknüpfung mit weiteren administrativen Datenquellen, etwa dem Betriebs- und Unternehmensregister des Bundesamts für Statistik oder dem Medizinal-

Unter den im ZSR für die Stichprobenziehung verfügbaren Informationen bestand die verfügbare Alternative damit nicht zwischen einer fehlerfreien und einer approximativen Messung der Praxisgrösse, sondern zwischen der Nutzung der verfügbaren K-Nummern und einem *Verzicht* auf die Praxisgrösse als Schichtungsmerkmal.

Vor diesem Hintergrund sind zwei Fragen voneinander zu unterscheiden.

- Zunächst ist die von Polynomics beschriebene Wirkungskette des postulierten **Verzerrungsmechanismus** methodisch einzuordnen; vgl. Kapitel 2.1.
- Anschliessend ist zu prüfen, welche **quantitative Bedeutung** eine allfällige Fehlklassifikation grosser Praxen für die geschätzte mittlere Behandlungsdauer überhaupt haben könnte; vgl. Kapitel 2.2.

## 2.1. Methodische Einordnung des von Polynomics postulierten Verzerrungsmechanismus

Polynomics argumentiert, dass eine unzureichende Berücksichtigung grosser Praxen zu einer systematischen Überschätzung der mittleren Behandlungsdauer führen könnte, sofern grosse Praxen tatsächlich kürzere Behandlungszeiten aufweisen:

„Sollten grosse Praxen tatsächlich kürzere Behandlungszeiten aufweisen und gleichzeitig in der Rahmenpopulation, der Stichprobe oder der Gewichtung zu wenig berücksichtigt worden sein, würden ihre kürzeren Zeiten zu wenig Gewicht erhalten. Die in der LeDa-Studie ausgewiesenen durchschnittlichen Behandlungszeiten könnten dadurch systematisch zu hoch ausfallen. Eine solche Verzerrung wäre für die Tarifbildung unmittelbar relevant.“<sup>8</sup>

Polynomics legt weder eine methodische Herleitung noch Literaturverweise vor, welche den postulierten Verzerrungsmechanismus näher begründen. Es ist daher zu prüfen, unter welchen Voraussetzungen die zitierte Argumentation zutrifft und welche Bedeutung dabei insbesondere der Schichtung, den Designgewichten und der Nachschichtung zukommt. Für diese Beurteilung sind zunächst zwei unterschiedliche Sachverhalte auseinanderzuhalten, auf die gesondert eingegangen werden muss.

### 2.1.1. Eine mögliche Fehlklassifikation innerhalb des Stichprobenrahmens

Soweit eine Fehlklassifikation lediglich die Zuordnung einer im Stichprobenrahmen enthaltenen Praxis zu einer Grössenklasse betrifft, führt sie bei einer Wahrscheinlichkeitsstichprobe nicht bereits als solche zu verzerrten designbasierten Schätzungen. Massgeblich ist vielmehr, dass die betreffenden Praxen im Stichprobenrahmen enthalten sind, eine positive Einschlusswahrscheinlichkeit besitzen und bei der Schätzung entsprechend den aus dem tatsächlich realisierten Stichprobenplan resultierenden Einschlusswahrscheinlichkeiten gewichtet werden.<sup>9</sup> Diese Voraussetzungen sind für die im ZSR enthaltenen und in die Stichprobenziehung einbezogenen Praxen der LeDa-Studie erfüllt. Davon zu unterscheiden ist eine mögliche Unterabdeckung der Zielpopulation durch den Stichprobenrahmen; diese wird in Abschnitt 2.1.2 gesondert behandelt.

Der von Polynomics postulierte Verzerrungsmechanismus greift daher in dieser Allgemeinheit nicht: Eine fehlerhafte Schichtzuweisung führt nicht bereits als solche zu einem falschen Designgewicht und begründet für sich genommen keine systematische Verzerrung der designbasierten Schätzung.

Eine approximative Klassifikation der Praxisgrösse ist gleichwohl methodisch relevant. Sie kann dazu führen, dass die durch die Schichtung angestrebte Homogenität innerhalb der Schichten geringer ausfällt und damit der Präzisionsgewinn der Schichtung, etwa in Form kleinerer Konfidenzintervalle, reduziert wird. Darüber hinaus kann sie die Wirksamkeit einer auf der Praxisgrösse beruhenden Nonresponse-Anpassung bzw. Nachschichtung beeinträchtigen.

---

beruferegister des Bundesamts für Gesundheit, nicht möglich.

<sup>8</sup> Polynomics, S. 5.

<sup>9</sup> Vgl. Dalenius, T./Ghosh, S. P., Some theory of sampling when the stratification is subject to error. *Technometrics*, 9 1967, Nr. 1, S. 1–13 zur theoretischen Herleitung sowie Mitani, A. A. et al., Survey design and analysis considerations when utilizing misclassified sampling strata. *BMC Medical Research Methodology*, 21 2021, Nr. 145, S. 1–13 für eine simulationsbasierte Untersuchung der Auswirkungen fehlerhafter Schichtzuordnungen.

Für die Beurteilung dieser Nachteile ist jedoch nicht ein hypothetisches Design mit fehlerfrei beobachteter Praxisgrösse der sachgerechte Vergleichsmassstab. Eine solche Information stand für die LeDa-Erhebung nicht zur Verfügung. Bei gegebenem Stichprobenrahmen bestand die unmittelbar verfügbare Alternative vielmehr darin, die vorhandene approximative Information zur Praxisgrösse zu verwenden oder auf die Praxisgrösse als Schichtungs- bzw. Anpassungsmerkmal zu verzichten. Auch eine fehlerbehaftete, aber informative Hilfsvariable kann gegenüber einem Verzicht auf diese Information weiterhin einen Präzisionsgewinn ermöglichen.<sup>10</sup> Entsprechendes gilt grundsätzlich für die Verwendung der Praxisgrösse bei der Nonresponse-Anpassung: Aus der Unvollkommenheit der Grössenklassifikation allein folgt weder, dass die Schichtung nach Praxisgrösse gegenüber einem Design ohne Grössenschichtung nachteilig war, noch dass eine Nonresponse-Anpassung ohne dieses Merkmal vorzuziehen gewesen wäre.

### 2.1.2. Eine mögliche Unterabdeckung der Zielpopulation

Von einer fehlerhaften Zuordnung der Praxisgrösse innerhalb des Stichprobenrahmens sind Abdeckungsfehler und Fehler bei der Bildung der Rahmeneinheiten zu unterscheiden. Eine Unterabdeckung läge etwa dann vor, wenn das ZSR Teile der Zielpopulation der zulasten der OKP abrechnenden Leistungserbringer nicht erfassen würde. Dies wäre ein Problem der Vollständigkeit des Stichprobenrahmens (ZSR) und nicht der Stichprobenziehung der LeDa-Studie. Die Qualität des ZSR wird nicht beanstandet. Davon wiederum zu unterscheiden wären Fehler bei der Bildung der Praxiseinheiten. Für eine Unterabdeckung des ZSR hinsichtlich der zulasten der OKP abrechnenden Leistungserbringer liegen aus unserer Sicht keine Anhaltspunkte vor. Auch aus den von Polynomics angeführten Vergleichen lässt sich ein Fehler bei der Bildung der Praxiseinheiten nicht unmittelbar ableiten, da sich die verwendeten Einheiten und Definitionen zwischen den Datenquellen unterscheiden können.

### 2.1.3. Einordnung und Zusammenfassung

Für die Beurteilung sind mehrere unterschiedliche Sachverhalte auseinanderzuhalten: die Zuordnung der Praxisgrösse innerhalb des Stichprobenrahmens, mögliche Abdeckungs- oder Einheitenfehler des Stichprobenrahmens sowie die Berücksichtigung der Gewichtung.

Eine fehlerhafte Zuordnung von Praxen zu Grössenklassen innerhalb des Stichprobenrahmens führt nicht bereits als solche zu einer systematischen Verzerrung der designbasierten Schätzung, kann jedoch den mit der Schichtung oder der Nonresponse-Anpassung angestrebten Effizienzgewinn vermindern. Abdeckungs- oder Einheitenfehler des Stichprobenrahmens könnten demgegenüber grundsätzlich zu verzerrten Schätzungen führen. Für eine Unterabdeckung des ZSR hinsichtlich der zulasten der OKP abrechnenden Leistungserbringer liegen aus unserer Sicht keine Anhaltspunkte vor; auch aus den von Polynomics angeführten Vergleichen lässt sich ein Fehler bei der Bildung der Praxiseinheiten nicht unmittelbar ableiten.

Soweit Polynomics eine systematische Verzerrung aus einer fehlerhaften Zuordnung von Praxen zu Grössenklassen innerhalb des Stichprobenrahmens herleitet, ist dieser Mechanismus daher sachlich nicht begründet. Eine solche Fehlklassifikation begründet für sich genommen keine systematische Verzerrung der designbasierten Schätzung.

## 2.2. Sensitivitätsanalyse: Quantitative Einordnung des von Polynomics beschriebenen Effekts

Wie im vorangehenden Abschnitt gezeigt, folgt aus einer fehlerhaften Zuordnung von Praxen zu Grössenklassen innerhalb des Stichprobenrahmens für sich genommen keine systematische Verzerrung der designbasierten Schätzung. Die folgende Sensitivitätsanalyse ist daher nicht als Schätzung einer dadurch verursachten Verzerrung zu verstehen, sondern als kontrafaktischer Stress-Test der von Polynomics beschriebenen Konstellation eines höheren Anteils grosser Praxen bei zugleich kürzeren Behandlungsdauern.

<sup>10</sup> Dalenius und Ghosh vergleichen eine Schichtung auf Grundlage fehlerbehafteter Klassifikationsinformationen ausdrücklich mit einem ungeschichteten Stichprobenverfahren und zeigen, dass die fehlerbehaftete Schichtung je nach Ausmass der Fehlklassifikation und der Stichprobenallokation effizienter oder weniger effizient sein kann; vgl. Dalenius/Ghosh, S. 9–11.

Polynomics schätzt auf Grundlage eigener Auswertungen des GLN-Registers den Anteil grosser Praxisstandorte, d. h. Standorte mit mind. vier Physiotherapeut/innen, auf rund 9%.<sup>11</sup> Dieser Anteil liegt 6.65 Prozentpunkte über dem in der LeDa-Studie ausgewiesenen Anteil grosser Praxen von 2.35%.<sup>12</sup> Die 9% werden im Folgenden deshalb nicht als Schätzung des tatsächlichen Anteils grosser Praxen, sondern als alternatives Szenario verwendet. Untersucht wird:

Wie stark würde sich die geschätzte durchschnittliche Behandlungsdauer verändern, wenn der Anteil grosser Praxen in der Zielpopulation tatsächlich höher wäre als in der LeDa-Rahmenpopulation ausgewiesen und grosse Praxen zugleich kürzere Behandlungszeiten aufwiesen?

Die Analyse dient somit weder einer nachträglichen "Korrektur" des LeDa-Schätzers noch der Quantifizierung einer durch fehlerhafte Schichtzuweisungen verursachten Verzerrung. Sie zeigt vielmehr, wie stark sich der Gesamtmittelwert unter den von Polynomics postulierten Annahmen verändern würde.

Wird der von Polynomics abgeleitete Anteil von rund 9% als alternatives Szenario verwendet, ergäben sich bei unveränderter Gesamtzahl von 5'368 Praxen rund 483 grosse Praxen anstelle der 126 in der LeDa-Rahmenpopulation. Dies entspricht rund 357 zusätzlichen grossen Praxen bzw. einer Erhöhung ihres Anteils um 6.65 Prozentpunkte. Wird vereinfachend unterstellt, dass diese zusätzlichen grossen Praxen vollständig zulasten der Kategorie der kleinen Praxen gehen, ergibt sich ein kontrafaktischer Mittelwert der Behandlungsdauer von rund 34.15 Minuten:

$$\text{Mittelwert}_{\text{kontra}} = 0.767 \cdot 34.41 + 0.143 \cdot 33.49 + 0.090 \cdot 32.95 \approx 34.15 \text{ Minuten.}$$

Gegenüber dem entsprechenden LeDa-Mittelwert von rund 34.24 Minuten ergibt sich damit eine Reduktion um rund 0.10 Minuten bzw. 5.8 Sekunden.<sup>13</sup> Im Verhältnis zur durchschnittlichen Behandlungsdauer fällt diese Veränderung mit weniger als 0.3% sehr gering aus.

Über die von Polynomics formulierten Einwände hinaus erscheint es sinnvoll, die Sensitivitätsanalyse um weitergehende Stressszenarien zu ergänzen. Damit wird geprüft, wie empfindlich die geschätzte mittlere Behandlungsdauer gegenüber zusätzlichen, für die LeDa-Schätzung bewusst ungünstigen Annahmen reagiert.

Neben einem auf 9% erhöhten Anteil grosser Praxen wird hierzu für sämtliche grossen Praxen eine gegenüber dem empirisch beobachteten Wert von 32.95 Minuten zunehmend niedrigere mittlere Behandlungsdauer unterstellt. Die Stressszenarien dienen damit einer weitergehenden wissenschaftlichen Einordnung der *möglichen* Grössenordnung des diskutierten Effekts. Die Einzelheiten der Berechnungen finden sich in Anhang A.

Die Ergebnisse in Anhang A zeigen, dass die geschätzte mittlere Behandlungsdauer selbst unter diesen bewusst ungünstigen Annahmen lediglich um rund 10 bis 22 Sekunden gegenüber dem LeDa-Schätzwert zurückgeht.

Die Stressszenarien zeigen damit, dass unter den betrachteten, bewusst ungünstig gewählten Annahmen nur eine begrenzte Veränderung des Gesamtmittelwerts resultiert. Diese Abweichungen sind nicht als Schätzung eines durch fehlerhafte Schichtzuweisungen verursachten Bias zu interpretieren, sondern ergeben sich aus der kontrafaktisch unterstellten Verteilung der Praxisgrössen und den zusätzlich angenommenen niedrigeren Behandlungsdauern grosser Praxen. Unter den betrachteten Szenarien ist die quantitative Bedeutung der von Polynomics thematisierten möglichen Abweichung daher gering.

---

<sup>11</sup> Polynomics, S. 13.

<sup>12</sup> Polynomics hält ausdrücklich fest, dass dieser Vergleich für sich genommen keine fehlerhafte Klassifikation belegt, da sich Grundgesamtheit, Erfassungszeitpunkt und Definition des Praxisstandorts unterscheiden können; vgl. Polynomics, S. 14.

<sup>13</sup> Die Annahme, dass die zusätzlich als gross eingestuften Praxen vollständig der Kategorie der kleinen Praxen entnommen werden, ist bewusst konservativ gewählt. Würden die zusätzlich grossen Praxen teilweise oder vollständig aus der mittleren Grössenklasse stammen, fiel die Veränderung entsprechend geringer aus.

### 3. Vermeintliche Mängel bei der statistischen Auswertung

Die von Polynomics festgestellte Diskrepanz<sup>14</sup> hinsichtlich der statistischen Signifikanz lässt sich anhand der finalen Fassung des LeDa-Schlussberichts nicht nachvollziehen. Polynomics bestätigte, dass der Begutachtung nicht die finale Fassung des LeDa-Schlussberichts vom 22. Oktober 2025 zugrunde lag.

In der finalen Fassung wird der Unterschied der mittleren Behandlungsdauer zwischen selbständig und angestellt tätigen Physiotherapeut/innen als statistisch signifikant ausgewiesen.<sup>15</sup>

#### 3.1. Die Nachrechnung von Polynomics bildet die designbasierte Inferenz nicht ab

Polynomics nimmt anhand der im LeDa-Bericht ausgewiesenen Mittelwerte, Standardfehler und nominellen Gruppenumfänge eine eigene Nachrechnung vor. Dabei werden die designbasierten Standardfehler mit den nominellen Gruppenumfängen von 305 bzw. 30 Beobachtungen im Rahmen einer klassischen Welch-Satterthwaite-Approximation kombiniert.<sup>16</sup> Polynomics weist selbst darauf hin, dass diese Nachrechnung voraussetzt, dass die ausgewiesenen Standardfehler unmittelbar für einen Welch-Test verwendet werden können, und dass das Ergebnis bei Berücksichtigung weiterer Aspekte des Stichprobendesigns abweichen kann.<sup>17</sup>

Die im LeDa-Bericht verwendete Inferenz beruht demgegenüber auf einem designbasierten t-Test unter Berücksichtigung von Schichtung und Gewichtung. Dieser ergibt  $|t| = 2.265$  bei 310 Freiheitsgraden und  $p = 0.024$ . Die von Polynomics vorgenommene Nachrechnung stellt daher keine korrekte Rekonstruktion dieses designbasierten Tests dar.

Die klassische Welch-Satterthwaite-Approximation bestimmt die Freiheitsgrade auf Grundlage der gruppenspezifischen Stichprobenumfänge und Varianzschätzungen unter einer Varianzstruktur, wie sie unabhängigen einfachen Zufallsstichproben zugrunde liegt.<sup>18</sup> Diese Voraussetzung ist bei der LeDa-Erhebung nicht gegeben.<sup>19</sup> Insbesondere können designbasierte Standardfehler nicht mit den nominellen Gruppenumfängen zu einer klassischen Welch-Satterthwaite-Approximation kombiniert werden.<sup>20</sup>

Soweit sich Unterschiede zwischen der Nachrechnung von Polynomics und der designbasierten Auswertung ergeben, sind sie daher nicht als Hinweis auf eine fehlerhafte Auswertung der LeDa-Studie durch Ecoplan/FHNW zu interpretieren. Die von Polynomics verwendete Welch-Satterthwaite-Approximation bildet die im LeDa-Bericht verwendete designbasierte Inferenz nicht korrekt ab.

#### 3.2. Dokumentation

Die Dokumentation der Datenplausibilisierung und -aufbereitung im LeDa-Bericht ist ausbaufähig. Aus der fehlenden Dokumentation lässt sich, wie Polynomics selbst festhält,<sup>21</sup> jedoch nicht ableiten, dass entsprechende Prüfungen unterblieben sind.

Im Rahmen der LeDa-Studie wurden umfangreiche Plausibilisierungs- und Datenprüfungen durchgeführt, die im Schlussbericht nur unzureichend dokumentiert wurden. Die entsprechenden Arbeitsschritte werden deshalb in Anhang A ergänzend dargestellt.

---

<sup>14</sup> Polynomics, S. 16–17 und Anhang A.

<sup>15</sup> Ecoplan/FHNW, Tabelle 12, S. 54.

<sup>16</sup> Vgl. Polynomics, Anhang A, S. 27.

<sup>17</sup> Polynomics, S. 17.

<sup>18</sup> Vgl. bspw. Bickel, P. J./Doksum, K. A., *Mathematical Statistics: Basic Ideas and Selected Topics*. Band 1, 2. Auflage. Boca Raton (FL): CRC Press, 2015, S. 264–266.

<sup>19</sup> Die LeDa-Erhebungen sind geschichtete Zufallsstichproben mit disproportionaler Stichprobenallokation; vgl. Ecoplan/FHNW, S. 12–13 und Anhang A.

<sup>20</sup> Die Bestimmung der Freiheitsgrade sowie weitere Aspekte des Testens im Rahmen designbasierter Inferenz werden bspw. bei Graubard, B. I./Korn, E. L., *Analysis of Health Surveys*. New York: John Wiley & Sons, 1999, S. 75–76 und S. 193–199 behandelt. Für die inferenzstatistischen Auswertungen der LeDa-Studie wurde die Funktionalität des R-Pakets *survey* verwendet; vgl. Ecoplan/FHNW, S. 44. Zur konkreten Implementierung vgl. Lumley, T. *survey: Analysis of Complex Survey Samples*, R package version 4.5, 2026, URL: <https://CRAN.R-project.org/package=survey>; zum methodischen Hintergrund des Pakets vgl. Lumley, Kapitel 2–7.

<sup>21</sup> Vgl. Polynomics, S. 17.

Zu präzisieren ist hingegen die Aussage von Polynomics, eine "abschliessende Bewertung der Datenqualität" sei allein anhand zusätzlicher Angaben der Autoren möglich. Solche Angaben erlauben insbesondere eine Beurteilung der vorgenommenen Qualitätssicherung und ihrer methodischen Angemessenheit. Eine abschliessende Beurteilung der tatsächlich erreichten Datenqualität lässt sich daraus jedoch nicht ableiten, da die Datenaufbereitung lediglich einen Teil der relevanten Qualitätsdimensionen des Total-Survey-Error-Ansatzes erfasst.<sup>22</sup>

## A. Anhang

### A.1. Stresstest: Weitere kontrafaktische Szenarien

Über die von Polynomics formulierten Einwände *hinaus* wird die Sensitivitätsanalyse in Kapitel 2.2 um weitere Stressszenarien ergänzt. Damit wird untersucht, wie stark sich die geschätzte mittlere Behandlungsdauer unter zusätzlichen, für die LeDa-Schätzung bewusst ungünstigen Annahmen verändern würde. Die Analyse wurde von Polynomics nicht verlangt und dient einer weitergehenden wissenschaftlichen Einordnung der möglichen Grössenordnung der von Polynomics thematisierten Abweichung. Sie ist nicht als Schätzung einer durch fehlerhafte Schichtzuweisungen verursachten Verzerrung zu verstehen.

In allen Stressszenarien wird der Anteil grosser Praxen auf 9% erhöht. Wie in Kapitel 2 wird dabei konservativ unterstellt, dass die im Szenario zusätzlich der Kategorie der grossen Praxen zugerechneten Praxen vollständig zulasten der Kategorie der kleinen Praxen gehen. Zusätzlich wird für sämtliche grossen Praxen eine gegenüber dem empirisch beobachteten Wert von 32.95 Minuten zunehmend niedrigere mittlere Behandlungsdauer angenommen; siehe Tabelle 2. Die Szenarien sind als Sensitivitätsanalyse ausgewählter, bewusst ungünstiger Konstellationen zu verstehen und nicht als Bestimmung einer maximal möglichen Abweichung.

Szenario A entspricht dem in Kapitel 2 behandelten Fall. In den Szenarien B–D wird die mittlere Behandlungsdauer grosser Praxen schrittweise auf 32, 31 bzw. 30 Minuten abgesenkt. Im bewusst ungünstig gewählten Szenario D beträgt die Veränderung gegenüber dem LeDa-Schätzwert rund 22 Sekunden. Unter den betrachteten kontrafaktischen Szenarien ist die quantitative Bedeutung der von Polynomics thematisierten möglichen Veränderung der mittleren Behandlungsdauer daher gering.

| Szenario | Annahme zum Anteil grosser Praxen | Annahme zur Behandlungsdauer grosser Praxen | Veränderung ggü. LeDa |
|----------|-----------------------------------|---------------------------------------------|-----------------------|
| Referenz | 2.35%                             | 32.95 Min.                                  | –                     |
| A        | 9.00%                             | 32.95 Min.                                  | - 5.8 Sekunden        |
| B        | 9.00%                             | 32.00 Min.                                  | - 10.9 Sekunden       |
| C        | 9.00%                             | 31.00 Min.                                  | - 16.4 Sekunden       |
| D        | 9.00%                             | 30.00 Min.                                  | - 21.8 Sekunden       |

**Tabelle 2:** Veränderungen der mittleren Behandlungsdauer gegenüber der LeDa-Studie unter Stressszenarien. Daten: LeDa-Studie, Selbstbeobachtung.

### A.2. Dokumentation zu Datenplausibilisierung

Im Rahmen der Datenaufbereitung wurden die erhobenen Angaben systematisch auf formale und inhaltliche Plausibilität geprüft. Dabei wurden insbesondere Wertebereiche und logische Konsistenzen kon-

<sup>22</sup> Vgl. Groves, R. M. et al., *Survey Methodology*. 2. Auflage. Hoboken (NJ): John Wiley & Sons, 2009, Kapitel 2 oder auch Biemer, P. P./Lyberg, L., *Introduction to Survey Quality*. Hoboken (NJ): John Wiley & Sons, 2003, Kapitel 1–2.

trolliert sowie fehlende und auffällige Angaben identifiziert und geprüft.<sup>23</sup>

Regelbasierte Plausibilitätsprüfungen können naturgemäss nur solche Auffälligkeiten erfassen, die durch die verwendeten Prüffregeln abgedeckt werden.<sup>24</sup> Ergänzend wurde deshalb eine ergebnisbezogene Robustheitsprüfung (outcome-based assessment) vorgenommen. Diese setzt nicht voraus, dass ungewöhnliche oder potenziell einflussreiche Beobachtungen vorgängig vollständig identifiziert werden. Hierzu wurden die konventionellen designbasierten Mittelwertschätzungen mit robusten Schätzungen auf Grundlage eines adaptiven *M*-Schätzers verglichen.<sup>25</sup> Die Ergebnisse unterscheiden sich nur geringfügig (vgl. Tabelle 3). Für kleine, mittlere und grosse Praxen betragen die robust geschätzten Mittelwerte 34.29, 33.33 bzw. 32.95 Minuten, gegenüber 34.41, 33.49 bzw. 32.95 Minuten bei der konventionellen Schätzung. Die maximale Abweichung beträgt damit 0.16 Minuten bzw. knapp 10 Sekunden. Die Ergebnisse liefern somit keine Hinweise darauf, dass die ausgewiesenen mittleren Behandlungsdauern wesentlich durch einzelne ungewöhnliche Beobachtungen (Ausreisser) beeinflusst werden.

| Praxisgrösse | Mittelwert | KI (unten) | KI (oben) |
|--------------|------------|------------|-----------|
| klein        | 34.29      | 33.29      | 35.28     |
| mittel       | 33.33      | 32.20      | 34.46     |
| gross        | 32.95      | 31.27      | 34.63     |

**Tabelle 3:** Robust geschätzte mittlere Behandlungsdauer (Minuten) nach Praxisgrösse und 95%-Konfidenzintervalle (MER-Schätzer); Daten: LeDa-Studie, Selbstbeobachtung

<sup>23</sup> Die Umsetzung erfolgte mit dem R Paket "validate"; vgl. van der Loo, M. P. J./de Jonge, E., Data Validation Infrastructure for R. Journal of Statistical Software, 97 2021, S. 1–31 und van der Loo, M./de Jonge, E., Statistical Data Cleaning with Applications in R. Hoboken (NJ): John Wiley & Sons, 2018, Kapitel 6.

<sup>24</sup> Vgl. bspw. van der Loo/de Jonge, 2018, Kapitel 8 und 9.

<sup>25</sup> Verwendet wurde der MER-Schätzer von Hulliger; vgl. Hulliger, B., Outlier Robust Horvitz–Thompson Estimators. Survey Methodology, 21 1995 sowie Hulliger, B., Horvitz–Thompson Estimators, Robustified. In Kotz, S. et al. (Hrsg.), Encyclopedia of Statistical Sciences. Band 5, 2. Auflage. Hoboken (NJ): John Wiley and Sons, 2006. Dabei handelt es sich um einen designbasierten adaptiven *M*-Schätzer. Für eine allgemeine Einführung in *M*-Schätzer vgl. bspw. Huber, P. J./Ronchetti, E., Robust Statistics. 2. Auflage. New York: John Wiley & Sons, 2009, Kapitel 3. Der MER-Schätzer ist im R-Paket "robsurvey" implementiert; vgl. Hulliger, B./Schoch, T., robsurvey: Robust Survey Statistics Estimation. 2026 (URL: <https://CRAN.R-project.org/package=robsurvey>).

## Literatur

- Bickel, P. J./Doksum, K. A.:** Mathematical Statistics: Basic Ideas and Selected Topics. Band 1, 2. Auflage. Boca Raton (FL): CRC Press, 2015
- Biemer, P. P./Lyberg, L.:** Introduction to Survey Quality. Hoboken (NJ): John Wiley & Sons, 2003
- Dalenius, T./Ghosh, S. P.:** Some theory of sampling when the stratification is subject to error. Technometrics, 9 1967, Nr. 1, 1–13
- Ecoplan/FHNW:** Studie zu den Leistungen in der Physiotherapie. Bern/Olten, 2025 – Studie im Auftrag von Physioswiss (Fassung vom 22. Okt. 2025)
- Graubard, B. I./Korn, E. L.:** Analysis of Health Surveys. New York: John Wiley & Sons, 1999
- Groves, R. M. et al.:** Survey Methodology. 2. Auflage. Hoboken (NJ): John Wiley & Sons, 2009
- Huber, P. J./Ronchetti, E.:** Robust Statistics. 2. Auflage. New York: John Wiley & Sons, 2009
- Hulliger, B.:** Outlier Robust Horvitz–Thompson Estimators. Survey Methodology, 21 1995, 79–87
- Hulliger, B.:** Horvitz–Thompson Estimators, Robustified. In **Kotz, S. et al. (Hrsg.):** Encyclopedia of Statistical Sciences. Band 5, 2. Auflage. Hoboken (NJ): John Wiley and Sons, 2006, 1–2
- Hulliger, B./Schoch, T.:** robsurvey: Robust Survey Statistics Estimation. 2026 (URL: <https://CRAN.R-project.org/package=robsurvey>), R package version 0.7–3
- Lumley, T.:** Complex Surveys: A Guide to Analysis Using R. Hoboken (NJ): John Wiley & Sons, 2010
- Mitani, A. A. et al.:** Survey design and analysis considerations when utilizing misclassified sampling strata. BMC Medical Research Methodology, 21 2021, Nr. 145, 1–13
- Polynomics:** Gesundheitsökonomische Prüfung der neuen KVG-Tarifstruktur für die Physiotherapie Studie zu Kosten und Leistungen in der Physiotherapie. Olten, 2026 – Gutachten im Auftrag von PROphysio (Fassung vom 10. Sept. 2026)
- van der Loo, M./de Jonge, E.:** Statistical Data Cleaning with Applications in R. Hoboken (NJ): John Wiley & Sons, 2018
- van der Loo, M. P. J./de Jonge, E.:** Data Validation Infrastructure for R. Journal of Statistical Software, 97 2021, 1–31