Tutorial

KARDIOTECHNIK 2023; 032(2):055–058
https://doi.org/10.47624/kt.032.PRGK3056

Statistik Teil 6: Methodenvergleiche

M. Kohl, F. Münch

Die KARDIOTECHNIK stellt in der Rubrik Tutorials relevante Methoden für wissenschaftliche Arbeiten zur klinischen Perfusion vor.

Fazitbox

Pro und Contra Methodenvergleiche:

Pro

- Bland-Altman-Diagramme stellen eine einfache Methode dar, um zwei Messmethoden miteinander zu vergleichen und sind klar einem einfachen Streudiagramm vorzuziehen.
- Mit Hilfe der Deming- und der Passing-Bablok-Regression lassen sich die Unterschiede zwischen zwei Messmethoden noch genauer analysieren.
- Bei Ausreißern, Abweichungen von der Normalverteilung oder sich ändernden Varianzen können Datentransformationen, nichtparametrische Bland-Altman-Diagramme und/oder die Passing-Bablok-Regression verwendet werden.

Contra

- Das klassische Bland-Altman-Diagramm sowie die Deming-Regression liefern nur bei konstanten Varianzen und normalverteilten Daten verlässliche Ergebnisse.
- Das klassische Bland-Altman-Diagramm und die Deming-Regression sind sehr sensitiv gegenüber Ausreißern.

Einführung

In der Biomedizin werden heute laufend neue Technologien eingeführt, insbesondere zur quantitativen Messung verschiedenster Parameter. Diese neuen Messmethoden müssen mit bestehenden (Goldstandard-) Verfahren verglichen werden, um ihre Einsetzbarkeit in der Praxis zu prüfen. Darüber hinaus sind viele bereits etablierte Messmethoden sehr sensitiv gegenüber den verwendeten Materialien. Hier sollte bei der Umstellung auf eine neue Materialcharge immer geprüft werden, ob sich dadurch die Messgenauigkeit des Verfahrens ändert. Im Folgenden werden wir die für diesen Zweck einsetzbaren Bland-Altman-Diagramme [1] sowie die Regressionsverfahren von Deming (1943) [2] und von Passing und Bablok (1983) [3] kurz vorstellen.

Bland-Altman-Diagramm

Das Bland-Altman-Diagramm wurde 1983 von Bland und Altman als eine einfache Möglichkeit zum Vergleich von zwei Messmethoden eingeführt [1]. Das Diagramm ist auch als Tukey Mittelwert Differenz-Diagramm (mean-difference plot) bekannt, welches dazu eingesetzt wird, eine Verschiebung zwischen zwei Verteilungen zu erkennen [4]. Im Rahmen der Analyse von genomischen Daten wird das Diagramm auch als MA-Plot bezeichnet [5]. Das (klassische) Bland-Altman-Diagramm ist ein Streudiagramm, bei dem auf der x-Achse der Mittelwert der beiden Messmethoden und auf der y-Achse die Differenz der beiden Messmethoden aufgetragen wird. Dies setzt demnach voraus, dass die gleiche Probe mit beiden Methoden gemessen wurde. Außerdem sollte man bereits vor der Durchführung der Messungen festlegen, welcher Unterschied zwischen den Methoden als (klinisch) relevant anzusehen ist.

Bei einem Streudiagramm, bei dem nur die Ergebnisse der beiden Verfahren dargestellt sind, ist es schwieriger, das genaue Ausmaß des Unterschiedes zu erkennen, vor allem bei einem großen Messbereich. In Abbildung 1 sind ein einfaches Streudiagramm und ein Bland-Altman-Diagramm zu sehen. Es handelt sich um Daten von 30 Patient:innen, die mit Hilfe extrakorporaler Zirkulation operiert wurden und bei denen die Activated Clotting Time (ACT) mit zwei gleichen Hemochron ACT-Geräten vor der Heparingabe gemessen wurde, wobei ein Messsystem mit einer LR-Küvette und ein Messsystem mit einer HR-Küvette ausgestattet war [6]. Die Daten der 30 Patient:innen sind auch in Tabelle 1 enthalten.

Wir können in Abbildung 1 von beiden Diagrammen ablesen, dass die ACT-Messung mit der LR-Küvette mit nur einer Ausnahme zu höheren Werten führte. Wir müssen daher davon ausgehen, dass es hier einen systematischen Unterschied (Bias) zwischen den beiden Messmethoden gibt. Während bei dem Bland-Altman-Diagramm die Daten das gesamte Diagramm ausfüllen, ist in einem solchen Fall bei einem einfachen Streudiagramm gewisser-

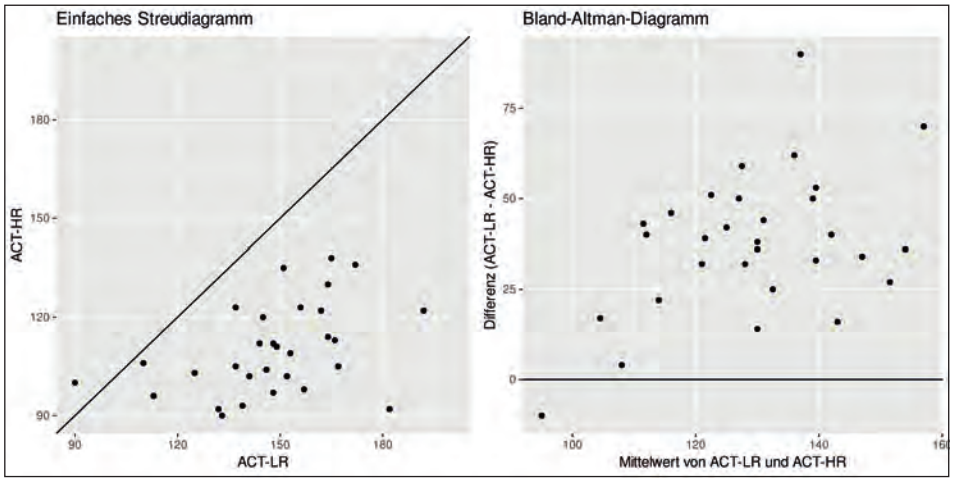

Abb. 1: Einfaches Streudiagramm und Bland-Altman-Diagramm von zwei ACT-Messungen mit gleichen Hemochron ACT-Geräten und unterschiedlichen Küvetten (LR vs. HR) vor der Heparingabe [6] (erstellt mit der Statistiksoftware R [7] und dem R Paket ggplot2 [8])

Prof. Dr. Matthias Kohl
Department of Medical and Life Sciences
Institute of Precision Medicine
Hochschule Furtwangen
Jakob-Kienzle-Str. 17,
78054 Villingen-Schwenningen (Germany)
E-Mail: kohl@hs-furtwangen.de
www.hs-furtwangen.de
www.life-data-science.org

maßen nur die Hälfte des Diagramms mit Daten gefüllt. Die Variabilität der Differenzen scheint über den gesamten Bereich der Mittelwerte der Messungen einigermaßen konstant zu sein. Sollte dies nicht der Fall sein, können die Daten zum Beispiel einer varianzstabilisierenden Transformation unterzogen werden. Oftmals eignet sich hierfür der Logarithmus.

Pat-ID	ACT-LR	ACT-HR
1	141	102
2	110	106
3	139	93
4	182	92
5	157	98
6	192	122
7	144	112
8	167	105
9	151	135
10	148	112
11	149	111
12	164	114
13	148	97
14	165	138
15	166	113
16	133	90
17	156	123
18	137	123
19	125	103
20	132	92
21	153	109
22	152	102
23	162	122
24	90	100
25	164	130
26	172	136
27	146	104
28	137	105
29	113	96
30	145	120

Tab. 1: Daten von 30 Patient:innen (Pat-ID 1–30), die mit Hilfe extrakorporaler Zirkulation operiert wurden und bei denen die Activated Clotting Time (ACT) mit zwei gleichen Hemochron ACT-Geräten vor der Heparingabe gemessen wurde, wobei ein Messsystem mit einer LR-Küvette (ACT-LR) und ein Messsystem mit einer HR-Küvette (ACT-HR) ausgestattet war [6]

Das einfache Streudiagramm in Abbildung 1 könnte dazu verleiten, für den Methodenvergleich die Pearson-Korrelation heranzuziehen. Die Pearson-Korrelation ist jedoch nicht für den Methodenvergleich geeignet, da damit nicht die Übereinstimmung, sondern die Stärke des linearen Zusammenhangs zwischen zwei Variablen gemessen wird [9]. Im Fall einer perfekten Übereinstimmung würden die Punkte der beiden Messungen alle auf der Winkelhalbierenden ($y = x$) liegen, während bei einem perfekten linearen Zusammenhang alle Punkte auf einer beliebigen Gerade ($y = ax + b$) liegen können. Entsprechend können Messungen mit einer schlechten Übereinstimmung trotzdem eine hohe Pearson-Korrelation aufweisen. Auch würde eine Änderung der Skalierung keinen Einfluss auf die Korrelation haben, während dies die Übereinstimmung deutlich verändern kann. Außerdem ist die Korrelation bei einem großen Wertebereich tendenziell höher als bei einem kleinen Wertebereich, während die Übereinstimmung unabhängig vom betrachteten Wertebereich sein sollte. Des Weiteren lässt sich die Übereinstimmung zwischen zwei Methoden auch nicht mit einem klassischen Signifikanztest untersuchen, da diese Tests generell darauf ausgelegt sind, Unterschiede zu finden [10]. Eine alternative Möglichkeit bestünde darin, sogenannte Äquivalenztests zu verwenden [11]. Wir werden uns hier jedoch auf Konfidenzintervalle [12] beschränken.

In einem Bland-Altman-Diagramm sollten zusätzlich der Mittelwert der Differenzen D, mit dem ein möglicher Bias zwischen den beiden Messungen identifiziert werden kann, sowie die untere und obere Übereinstimmungsgrenze (limit of agreement) eingezeichnet werden. Für die Festlegung der Übereinstimmungsgrenzen wird üblicherweise zusätzlich die Standardabweichung der Differenzen SD_D benötigt. Man definiert diese Grenzen meist als $D \pm 1{,}96 * SD_D$, wobei 1,96 gerade das 97,5 % Quantil der Standardnormalverteilung ist. Ausgehend von einer Normalverteilung sollten demnach theoretisch 95 % der Differenzen in diesem Intervall liegen; die untere und obere Übereinstimmungsgrenze sind gerade identisch zum 2,5 % und 97,5 % Quantil der Verteilung der Differenzen. In der Praxis kann dieser Anteil natürlich leicht variieren. Eine deutliche Abweichung vom erwarteten Anteil von 95 % kann durch schiefe Datenverteilungen oder Ausreißer verursacht werden. In einem solchen Fall ist eine normalisierende Datentransformation (auch hier eignet sich oftmals der Logarithmus), eine Verwerfung von Ausreißern oder die Verwendung alternativer nichtparametrischer oder robuster statistischer Methoden zu empfehlen [13,14]. Vor einer Verwerfung von Ausreißern sollte nach Möglichkeit die Ursache für die Ausreißer identifiziert und ausgeschlossen werden, dass es sich hierbei um systematische Abweichungen zwischen den beiden Methoden handelt.

In Abbildung 2 sind das klassische (parametrische) und ein nicht-parametrisches Bland-Altman-Diagramm dargestellt. Im klassischen Fall sind D und $D \pm 1{,}96 * SD_D$ dargestellt, wobei für SD_D hier die biasfreie Schätzung der Standardabweichung verwendet wurde [15]. Im nichtparametrischen Fall sind der Median sowie das 2,5 % und das 97,5 % (empirische) Quantil der Differenzen eingezeichnet.

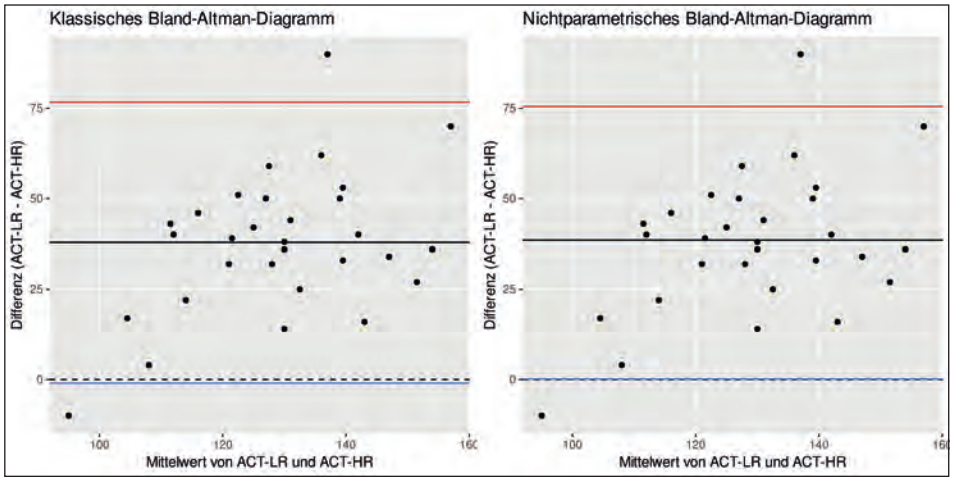

Abb. 2: Das klassische (parametrische) und ein nicht-parametrisches Bland-Altman-Diagramm der ACT-Daten [6] (erstellt mit der Statistiksoftware R [7] und dem R Paket MKinfer [16])

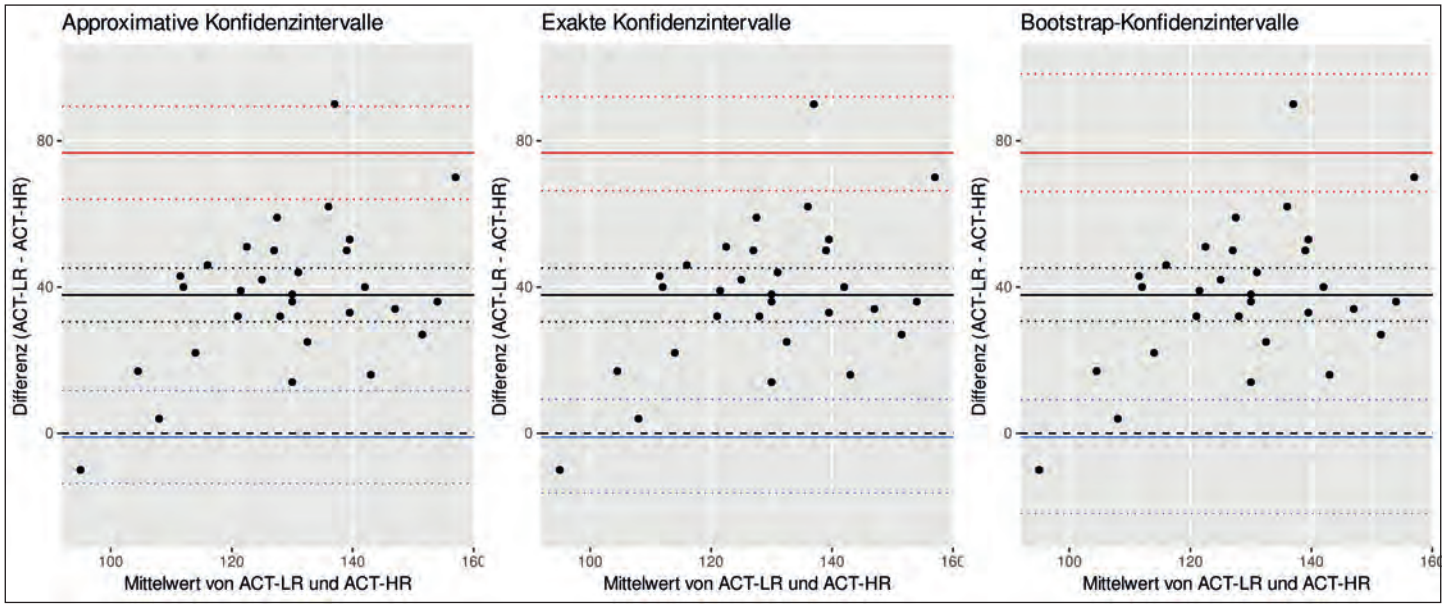

Abb. 3: Das klassische Bland-Altman-Diagramm der ACT-Daten [6] mit approximativen, exakten und Bootstrap-Konfidenzintervallen (erstellt mit der Statistiksoftware R [7] und dem R Paket MKinfer [16])

Die beiden Diagramme unterscheiden sich kaum, was dafür spricht, dass sich die Differenzen gut durch eine Normalverteilung beschreiben lassen. Dies wird auch dadurch bestätigt, dass in nur zwei Fällen die Differenzen außerhalb der Übereinstimmungsgrenzen liegen. Dies entspricht einer relativen Häufigkeit von 6,7 % und kann bei 30 Werten durchaus erwartet werden. In einem Bland-Altman-Diagramm sollten aber neben den geschätzten Werten für den Mittelwert und die Übereinstimmungsgrenzen auch Konfidenzintervalle für diese Werte angegeben werden [13,15]. Abbildung 3 zeigt die Erweiterung des klassischen Bland-Altman-Diagramms aus Abbildung 2 um entsprechende 95 %-Konfidenzintervalle (CI95). Es handelt sich um approximative [13], exakte [15] und Bootstrap-Konfidenzintervalle [12].

Während die approximativen Konfidenzintervalle alle symmetrisch sind, erhalten wir im Fall der exakten und der Bootstrap-Konfidenzintervalle (t-Methode, vgl. Supplement von [12]) im Fall der Übereinstimmungsgrenzen sichtbar asymmetrische Intervalle. Entsprechend sollten bei den Übereinstimmungsgrenzen besser die exakten Konfidenzintervalle verwendet werden [15]. Auch die Bootstrap-Konfidenzintervalle, die weniger Voraussetzungen benötigen, sind eine Alternative. Da aber recht extreme Quantile (2,5 % und 97,5 %) der Verteilung der Differenzen untersucht werden, sollten diese primär bei größeren Stichprobenumfängen zum Einsatz kommen. Der Bias beträgt 37,8 und ist signifikant von 0 verschieden, da keines der CI95 die 0 beinhaltet. Alle drei CI95 für den Bias sind nahezu identisch und reichen gerundet von 30,5 bis 45,2. Bei Mittelwerten der Messwerte im Bereich von ca. 100 bis 150 entspricht dies einer Abweichung im Bereich von 20 % oder mehr zwischen den beiden Messmethoden. Wir müssen daher davon ausgehen, dass der Bias zwischen den beiden Methoden auch (klinisch) relevant ist. Würde das CI95 des Bias innerhalb des vor den Messungen festgelegten (klinisch) relevanten Unterschiedes liegen, wäre das Ergebnis zwar signifikant, aber nicht (klinisch) relevant.

Die untere Übereinstimmungsgrenze liegt bei -1,0 (approximatives CI95: -13,7–11,7; exaktes CI95: -16,4–9,3; Bootstrap-CI95: -22,0–9,1), die obere Übereinstimmungsgrenze bei 76,7 (approximatives CI95: 64,0–89,4; exaktes CI95: 66,3–92,0; Bootstrap-CI95: 65,8–98,2). Die Übereinstimmungsgrenzen liegen demnach recht weit auseinander, weshalb wir nicht nur von einem signifikanten und (klinisch) relevanten Bias, sondern auch von einer recht großen Schwankungsbreite für die Unterschiede zwischen den Messungen ausgehen müssen. Kwapil et al. [6] vermuten, dass unterschiedliche Aktivatoren für diese recht deutlichen Unterschiede zwischen LR- und HR-Küvetten verantwortlich sind. Falls der Bias nicht signifikant bzw. zumindest nicht (klinisch) relevant ist, können die Übereinstimmungsgrenzen dazu herangezogen werden, um festzustellen, ob es trotzdem zu klinisch relevanten Unterschieden zwischen den beiden Messmethoden kommen kann.

Deming-Regression

Wie wir oben festgehalten haben, entspricht eine perfekte Übereinstimmung zwischen zwei Messmethoden im einfachen Streudiagramm gerade der Geraden y = x. Dies bedeutet, dass a = 1 und b = 0 für y = ax + b gelten muss. Der Achsenabschnitt steht in diesem Fall demnach für eine konstante Verschiebung und die Steigung für einen proportionalen Unterschied zwischen den beiden Messungen. Da die Messungen von beiden Methoden (also x und y) mit zufälligen Fehlern behaftet sind, eignet sich eine einfache lineare Regression jedoch nicht, um dies zu untersuchen, da hier nur zufällige Fehler in den y-Werten zugelassen sind; es kann zu irreführenden Ergebnissen führen. Stattdessen kann die Deming-Regression [17] verwendet werden, welche von Adcock 1878 eingeführt wurde [18]. Im Fall, dass die Varianzen für beide Messungen gleich sind, entspricht die Deming-Regression gerade der orthogonalen Regression, bei der die Summe der Quadrate der senkrechten Abstände von der Regressionsgeraden minimiert wird. In Abbildung 4 sind die beiden ACT-Messungen zusammen mit den Geraden der linearen und der Deming-Regression dargestellt.

Die beiden Regressionsgeraden unterscheiden sich sichtbar, insbesondere auch von der Geraden y = x. Im Fall der linearen Regression erhalten wir einen Achsenabschnitt von 68,8 (CI95: 34,9 – 107,7) und eine Steigung von 0,28 (CI95: 0,05–0,51), im Fall der Deming-Regression einen Achsenabschnitt von 55,4 (CI95: -5,6–106,4) und eine Steigung von 0,40 (CI95: 0,02–0,79). Dies bedeutet insbesondere, dass wir keine signifikante konstante Verschiebung, aber einen signifikanten proportionalen Unterschied zwischen den beiden Messun-

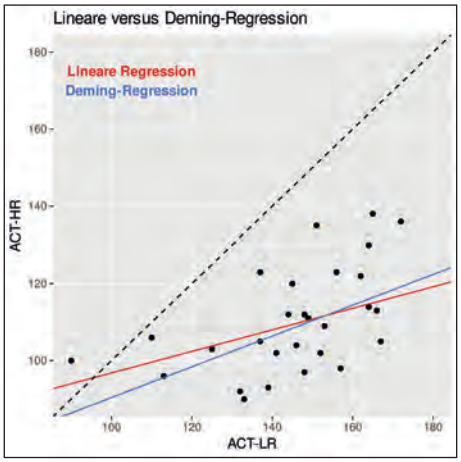

Abb. 4: Lineare und Deming-Regression für die ACT-Daten [6] (erstellt mit der Statistiksoftware R [7] und den R Paketen ggplot2 [8] und deming [19])

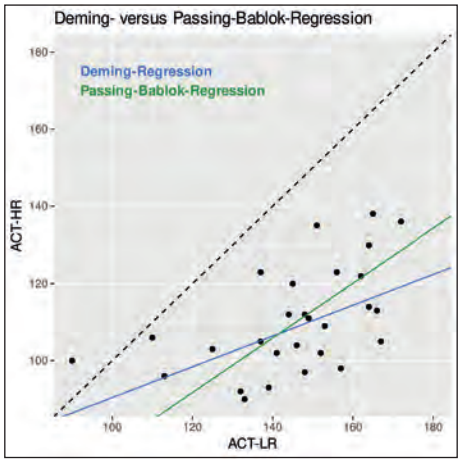

Abb. 5: Deming- versus Passing-Bablok-Regression für die ACT-Daten [6] (erstellt mit der Statistiksoftware R [7] und den R Paketen ggplot2 [8] und deming [19])

gen erhalten. Insofern bestätigt auch diese Analyse, dass es signifikante Unterschiede zwischen den beiden ACT-Messmethoden gibt. Das Ausmaß und damit die (klinische) Relevanz des Unterschiedes kann aber einfacher aus den Bland-Altman-Diagrammen in Abbildung 3 abgelesen werden.

Passing-Bablok-Regression

Da sowohl das klassische Bland-Altman-Diagramm als auch die Deming-Regression sensitiv gegenüber Ausreißern sind, betrachten wir abschließend die Passing-Bablok-Regression [20], welche von der Rangkorrelation nach Kendall (Kendalls τ) [9] abgeleitet werden kann. Da Rangkorrelationen robust gegenüber Ausreißern sind [9], gilt dies folglich auch für die Passing-Bablok-Regression [14]. In Abbildung 5 vergleichen wir die Deming- mit der Passing-Bablok-Regression für die ACT-Daten [6], wobei die skaleninvariante Variante der Passing-Bablok-Regression mit Bootstrap-Konfidenzintervallen berechnet wurde [21,19].

Die Ergebnisse der beiden Regressionsverfahren unterscheiden sich deutlich. Im Fall der Passing-Bablok-Regression erhalten wir einen Achsenabschnitt von 6,5 (CI95: -20,4–10,9) und eine Steigung von 0,71 (CI95: 0,68–0,89). Damit ergibt sich auch bei der Passing-Bablok-Regression keine signifikante konstante Verschiebung, aber ein signifikanter proportionaler Unterschied, wobei die Steigung deutlich näher zu 1 liegt als im Fall der Deming-Regression. Aufgrund der recht deutlichen Unterschiede bei den Ergebnissen und der fehlenden Robustheit der Deming-Regression sollten in diesem Fall besser die Ergebnisse der Passing-Bablok-Regression für den Methodenvergleich herangezogen werden.

Zusammenfassung

Das Bland-Altman-Diagramm stellt eine einfache Methode dar, um zwei Messmethoden miteinander zu vergleichen, und ist klar einem einfachen Streudiagramm vorzuziehen. Der Einsatzbereich des klassischen Bland-Altman-Diagramms lässt sich zudem durch Variablentransformationen oder den Einsatz von robusten oder nichtparametrischen Verfahren noch deutlich erweitern. Die Regressionsverfahren von Deming sowie von Passing und Bablok können zusätzlich dazu herangezogen werden, um die Übereinstimmung bzw. Abweichungen von der Übereinstimmung genauer zu analysieren. Durch ihre Robustheit ist die Passing-Bablok-Regression in den meisten Fällen der Deming-Regression vorzuziehen.

LITERATUR

1. Altman DG, Bland JM. Measurement in Medicine: the Analysis of Method Comparison Studies. The Statistician 1983; 32:307-317.

2. Deming WE. Statistical adjustment of data. Wiley, NY 1943. (Dover Publications edition, 1985).

3. Passing H, Bablok W. A New Biometrical Procedure for Testing the Equality of Measurements from Two Different Analytical Methods. J. Clin. Chem. Clin. Biochem 1983. 21:709-720.

4. Cleveland WS. Visualizing data. Murray Hill, N.J.: At & T Bell Laboratories 1993. 22-23.

5. Dudoit S, Yang YH, Callow MJ, Speed TP. Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. Stat. Sin 2002. 12(1):111-139

6. Kwapil N, Teske A, Einhaus F, Krajinovic L, Purbojo A, Dittrich S, Dewald O, Münch F. Aussagekraft einer Activated Clotting Time-Messung. Kardiotechnik 2022. 31(3):90-94.

7. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing 2022. Vienna, Austria. URL https://www.R-project.org/.

8. Wickham H. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York 2016.

9. Kohl M, Münch F. Statistik Teil 4: Korrelationen. Kardiotechnik 2022. 31(4):146-149.

10. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1986.327(8476):307-10.

11. Shieh G. Assessing Agreement Between Two Methods of Quantitative Measurements: Exact Test Procedure and Sample Size Calculation. Statistics in Biopharmaceutical Research 2020. 12(3): 352-359.

12. Kohl M, Münch F. Statistik Teil 3: Konfidenzintervalle. Kardiotechnik 2022. 31(3):95-98.

13. Bland JM, Altman DG. Measuring agreement in method comparison studies. Stat Methods Med Res 1999. 8:135-160.

14. Dufey F. Derivation of Passing-Bablok regression from Kendall's tau. Int J Biostat 2020. 16(2):20190157.

15. Shieh G. The appropriateness of Bland-Altman's approximate confidence intervals for limits of agreement. BMC Med Res Methodol 2018. 18:45.

16. Kohl M. MKinfer: Inferential Statistics. R package version 1.0. 2022

17. Deming WE. Statistical adjustment of data. Wiley, NY 1943 (Dover Publications edition, 1985).

18. Adcock RJ. A problem in least squares. The Analyst 1878. 5(2):53-54.

19. Therneau T. deming: Deming, Theil-Sen, Passing-Bablok and Total Least Squares Regression. R package version 1.4. 2018

20. Passing H, Bablok W. A new biometrical procedure for testing the equality of measurements from two different analytical methods. Application of linear regression procedures for method comparison studies in Clinical Chemistry, Part I. Journal of Clinical Chemistry and Clinical Biochemistry 1983. 21(11):709-720.

21. Bablok W, Passing H, Bender R, Schneider B. A general regression procedure for method transformations. Application of linear regression procedures for method comparison studies in Clinical Chemistry, Part III. Journal of Clinical Chemistry and Clinical Biochemistry 1988. 21. 26:783-790.