Stellungnahme zu

Effects of Integrated Care on Disease-Related Hospitalisation and Healthcare Costs in Patients with Diabetes, Cardiovascular Diseases and Respiratory Illness: A Propensity-Matched Cohort Study

Carola A. Huber, Oliver Reich, Mathias Früh, Thomas Rosemann

International Journal of Integrated Care, 16(2016)1:11, S. 1-18

Diese Stellungnahme schließt sich an die Darstellungen in „vemsreviews GEP Problem“[1] an. Es werden Kardinalfehler in der Methodik und in der angeblichen Verwendung von Mathematik in der Arbeit „Effects of Integrated Care …“ benannt. Teilweise zu „vemsreviews GEP Problem“ redundante Darstellungen sind in dieser Stellungnahme beabsichtigt.

Angefertigt durch: Dr. Walter Warmuth

erscheint in GwH

Zossen, 26.08.2016

[1] Verein Ethik und Medizin Schweiz, CH-4600 Olten

Inhaltsverzeichnis

HDoz Dr. rer. nat. habil. Walter Warmuth | Drosselgasse 5 | 15806 Zossen | **T** +49 3377 303621 | **F** +49 3377303622 | **E** ww@suljee.de

1. Vorbemerkung

Der Autor dieser Stellungnahme erklärt, dass für ihn das im Folgenden Dargestellte keinen Interessenkonflikten unterliegt. Die Darstellungen beruhen auf eigenen wissenschaftlichen Publikationen, für die er keine Zuwendungen aus öffentlichen Mitteln oder Projektfördermittel von Unternehmen oder Verbänden der privaten Wirtschaft erhalten hat. Der Autor war und ist Mitglied in wissenschaftlichen Fachgesellschaften, er hatte oder hat dort keine Ämter der Leitung oder Selbstverwaltung inne.

Praktisch sollte der Weg zu einer Studie die folgenden Schritte durchlaufen: Formulierung einer Zielsetzung, Datenauswahl, Datenaufbereitung (u. U. Erkunden von Botschaften in den Daten, u. U. Entwickeln von Datenintelligenz), Data Mining, Bewertung und Interpretation, Umsetzung, Kontrolle. Diese Schritte sind unabhängig von der Wahl eines Werkzeugs (eines Tools), ein Tool ersetzt auch höchstens den Strafarbeitsteil einer Datenaufbereitung.

2. Studiendesign

Die Autoren wählen für den Beleg ihres Zieles ein Effektivitätsmodell mit dem Fokus auf die Erklärung der Wirkung. Der Goldstandard für diese Art des Beleges wäre „Randomized Controlled Trials" zur Erkundung „signifikanter" Unterschiede hier zwischen einer Interventions- und Kontrollgruppe. Bei der nun jedoch vorgenommenen „Quasi-experimentellen Evaluation" wird die Randomisierung durch Auswahl- oder Matchingverfahren ersetzt. Das Auswahlverfahren muss schon etwas mit den Behandlungen, mit der Versorgung, …, die man miteinander vergleichen will, zu tun haben. Diese fachlich inhaltliche Beziehung kann nicht durch eine Deklaration von Beziehung bzw. durch irgendwelche Benennungen ersetzt werden. Dass die Ein- und Ausschlüsse in die Auswahl fachgerecht sind, muss begründet werden. Regressionskoeffizienten und Variablenausprägungen ergeben den Score Wert. Bei welcher „Distanz" zwischen den Score Werten eine Ähnlichkeit vermutet wird, ist zu benennen.

3. *Bildung einer Kontrollgruppe*

Es ist äußerst fatal, wenn eine Ähnlichkeit (vor der Intervention) zwischen der Interventionsgruppe und der Kontrollgruppe nicht hergestellt werden kann. Zu berücksichtigen sind alle Personen, die für die Gruppen in Frage kamen … auch die Verweigerer, auch die, die nach genauerer Betrachtung (z. B. Anamnese) „abgewählt" werden, die sich vergleichbar compliant zur Therapie verhalten, die vergleichbar adherent auf die Therapie reagieren … für alle diese Fälle hat sich ein „Zwilling" in der Kontrollgruppe zu finden. Die Autoren lassen hier jede Sorgfalt vermissen, mehr noch: die gebildete Kontrollgruppe ist falsch, sie nehmen sie hin … um ihre Zielstellung nicht zu gefährden? Wenn weitgehend Rosinen per ICM herausgepickt werden, dann sind die „Rückfälle" (Rezidiv) seltener. Vermiedene Aufnahmen führen zu u. U. später schwerwiegenden Hospitalisierungen. Diesem Vergleich halten die gebildeten Kontrollgruppen nicht stand.

HDoz Dr. rer. nat. habil. Walter Warmuth | Drosselgasse 5 | 15806 Zossen | **T** +49 3377 303621 | **F** +49 3377303622 | **E** ww@suljee.de

Es kann nicht beruhigen, dass Frau Lüthy vom Universitätshospital Zürich versichert, dass die Ergebnisse der Studie „nicht zur Anwendung" gelangen werden. Immerhin bescheinigt Frau Lüthy den Autoren, dass die Matched Pairs „sorgfältig ausgearbeitet" wurden und „von hoher Qualität" sind. Ob es sich bei dem Abgeben der Erklärung von Frau Lüthy „nur" um eine dolose oder doch schon kriminelle Handlung handelt, ist zu beurteilen. Ist eine Leiterin der Unternehmenskommunikation intellektuell und fachwissenschaftlich überhaupt zu einer solchen Bewertung fähig? Beunruhigt es nicht die ersten drei Autoren, eine solche Bescheinigung aus der Unternehmenskommunikation zu erhalten? Hat der vierte Autor keine Skrupel?

4. Bewertung des Vergleiches

Zur Bewertung eines Vergleiches kann aus gleichen Stichproben zu verschiedenen Zeitpunkten (Längstschnittmethode) oder auf verschiedene Stichproben zum gleichen Zeitpunkt (Querschnittsmethode) zurückgegriffen werden. Diese Modellierungsergebnisse sind als Varianten der Ergodenhypothese gleich, denn jedes dynamische System kehrt unvermeidbar in einen Zustand zurück, der seinem Anfangszustand sehr nahe kommt. Es ist vernünftig, sich in dem konkreten Fall der Querschnittsmethode zu bedienen. Das Auswahlverfahren darf hierbei nicht auf Variable abzielen, deren Unterschiedlichkeit bei der Wahl schon die Unterschiedlichkeit bei der Bewertung begründen.

5. Fatale Variable Zeit

Zudem ist die unüberwindliche Hürde – analog zur Unschärferelation von W. Heisenberg – der Konfundierung zu beachten, dass zwei oder mehr unabhängige Variable zwar gleichzeitig variieren, experimentell jedoch nicht isoliert werden können. Konfundierungstests führen die Autoren nicht durch, eine Unkonfundiertheit der von den Autoren benutzten Variablen wird nicht belegt, einer offensichtlichen Teilkonfundierung zwischen den Variablen wird durch Ausschluss bestimmter Variabler oder durch die Bildung von neuen Variablen in der Arbeit nicht begegnet.

Insbesondere ist die Zeitdauer ein gefährlicher Vermittler eines scheinbaren Zusammenhanges (im Durchschnitt längere Tage sind im Durchschnitt wärmer, die Dauer ist aber keine Ursache für die Temperatur; die Fallzeit ist auch keine Ursache für den zurückgelegten Weg, …). Die Autoren des Papiers passen hier gleich zwei Mal: Die Interventionsgruppe bevorzugt erstens kürzere Verweildauern, denn die Kontrollgruppe ist vom Versorgungsauftrag her auch für eine Akutversorgung, auf Dauer, Stabilität und Zukunftssicherheit angelegt. Zweitens verdeckt die Zeit als Effektivitätsmaß jede Frage nach dem „gleichen" Effekt (der wird dann einfach unterstellt). Und was ist mit der Qualität der Versorgung?

HDoz Dr. rer. nat. habil. Walter Warmuth | Drosselgasse 5 | 15806 Zossen | **T** +49 3377 303621 | **F** +49 3377303622 | **E** ww@suljee.de

6. Diagnose-Systeme

Bei der Klassifizierung nach ICD10 (Internationale statistische Klassifikation der Krankheiten und verwandter Gesundheitsprobleme 10. Revision) handelt es sich um ein statistisches Kodierungssystem der Krankheiten und verwandter Gesundheitsprobleme. ICD10 sind keine Kosten, keine Behandlungsdauern, Verweildauern, … zuordenbar. Das DRG-System (Diagnostic Related Groups) ist ein Patienten-Klassifikationssystem, es berücksichtigt die individuelle Schwere der Erkrankung und individuelle Komorbiditäten, es berücksichtigt eine Fallgruppen-Schwere und Fallgruppen-Komorbiditäten, … Das Abrechnungssystem ist behandlungsorientiert, also aufwandsorientiert. Die Verweilzeit des Patienten findet pauschal Berücksichtigung. Bei einem Unterschreiten der unteren Korridorgrenze kommt es zu einem Erlösabschlag, beim Überschreiten der oberen Korridorgrenze werden Zuschläge erhoben. Mit den Pauschalen erhalten die Leistungserbringer eine Sicherheit, die sie als Akutversorger ohne eine besondere Spezialisierung für die Erfüllung ihrer Aufgaben benötigen. Bettenordnungen, Qualifizierung; Ausbildung, Anstellung und Bindung, … sind vorzuhalten.

Die Verweilzeit kann überhaupt nicht als Bewertungsmaßstab genutzt werden, weil sie in dem DRG-System keine Steuer- oder Kostengröße ist (zum Vergleich zwischen nach DRG abrechnenden Einrichtungen ist die Verweilzeit bedingt für die Suche nach einer Strukturoptimierung geeignet, zumindest wäre ein Nachdenken darüber sinnvoll).

7. Verschiedene Informationen

Es gibt viele Merkmale der Verschiedenheit. Verschieden sind zum Beispiel die Aufgreifkriterien, die die Diagnose betreffen. Eine objektive medizinische Diagnose würde sich auf Merkmale, die objektiv bestimmbar sind, gründen. Doch nicht bei jedem körperlichen Krankheitszustand können objektivierbare Merkmale bestimmt werden. Eine Therapie beginnt ein Arzt, wenn er auf der Basis seiner Erfahrung unter Kenntnis der Anamnese und für ihn wichtigen Voruntersuchungen hinreichend „viel" weiß, wenn er seine Wissenswahrscheinlichkeit hoch einschätzt.

Bei den ICD10 handelt es sich bestenfalls um statistische Diagnosen (als Hauptdiagnosen, Nebendiagnosen; Überweisungsdiagnosen; Einlieferungsdiagnosen, Aufnahmediagnosen), sie sind nicht therapiebezogen, sie haben keine Behandlungskosten, sie haben keine Behandlungszeiten, sie orientieren sich weder am Individuum (z. B. Anamnese) noch an mit der statistischen Kodierung in der Gruppe verbundenen üblichen Therapien, Therapien Standards, Komorbiditäten, … Die Auswahl für die Interventionsgruppe nach ICD10 ist stark durch die Vorstellung geprägt, im Rahmen bestimmter Verweilzeiten einen „fixen" Plan von Therapie zu realisieren.

Bei den DRG handelt es sich um Diagnosen, die die „mittlere" Behandlung in den Mittelpunkt stellen. Die nach DRG abrechnenden Einrichtungen sind in der Regel auch Akutversorger. Es kann schon sein, dass sie ihre Abrechnung nach dem höheren Erlös orientieren. Z. B. kann ein Hospital ohne Stroke Unit den Sturz eines Patienten aus dem Bett nicht als Schlaganfall abrechnen, erleidet ein Patient hingegen nach einer Hüftersatzoperation einen Schlaganfall (weil z. B. in der Anamnese ein früherer Schlaganfall nicht berücksichtigt wurde) und eine Stroke Unit ist vorhanden, dann verzichtet das Hos-

HDoz Dr. rer. nat. habil. Walter Warmuth | Drosselgasse 5 | 15806 Zossen | **T** +49 3377 303621 | **F** +49 3377303622 | **E** ww@suljee.de

pital auf die Abrechnung der zwar teuren Hüftersatzoperation, denn der Erlös aus dem Schlaganfall fällt als Fallpauschale höher aus – beides wäre nach dem DRG-System nicht abrechenbar. Wirkliche Verweilzeiten im Hospital sind nur für einen Vergleich mit Einrichtungen geeignet, die ebenfalls nach dem DRG-System abrechnen. Denn für diese Einrichtungen spielen die Bettenauslastung, spielt die Beschäftigung des Pflegepersonals, spielen verfügbare Kapazitäten ärztlicher Kunst … jetzt und in der Zukunft eine planerische Rolle. Hier geht es um ein soziales System, es geht um die Absicherung einer Zusage des Staates, es geht um eine Daseinsvorsorge für die Bürger.

8. Modellwahl

Der Score Wert begründet sich nicht nur auf Variablenausprägungen, sondern auch auf Regressionskoeffizienten. Beim Propensity Score Matching wird der Score oft mit Hilfe der Logistischen Regression gebildet. Dieser Score ist hervorragend geeignet, eine kleine Menge besonders guter Fälle aus vielen Fällen auszuwählen. Bei der Ansprache von besonders kaufwilligen Personen ist das bedenkenlos anwendbar. Einer Person einen Kauf nicht angeboten zu haben ist vertretbar, was ist aber beim Vorenthalten einer Therapie? Da es sich jedoch bei ICM und SCM um völlig unterschiedliche Herangehensweisen an die Versorgung handelt (unterschiedliche Diagnosesysteme, Spezialversorgung vs. Akutversorgung, Abrechnung nach Selektivaufwand vs. Fallpauschalen, …) sind auch nur unterschiedliche Regressionen „rechenbar". Die Skalen aus unterschiedlichen Logistischen Regressionen sind nicht miteinander vergleichbar, so wie es etwa Prozentangaben der Temperaturveränderung bei °K und °C auch nicht sind. Mathematisch verbrämt werden also sowohl auf die Daten der Interventionsgruppe als auch auf die der Kontrollgruppe Modelle der linearen Regression angewandt. Dieses Modellieren bedeutet:

1. Die abhängige Variable wird als normalverteilt modelliert.

2. Für die abhängige normalverteilte Variable wird als Modell eine lineare Funktion unabhängiger Variabler und ein zufälliger Fehler, der für alle anderen „Einflüsse" steht, gewählt.

3. Die zufälligen Fehler müssen sich im Mittel aufheben.

4. Die unabhängigen Variablen sind nicht zufällig, eine neue Beobachtung zeigt in Bezug auf diese Variablen immer das gleiche Ergebnis.

5. Die unabhängigen Variablen sind linear unabhängig.

6. Schwankungen sind homogen, die Streuungen sind gleich.

7. Schwankungen sind nicht autokorreliert, Schwankungen verschiedener Beobachtungen sind unkorreliert.

Eine Abweichung von der Modellannahme der Normalverteilung kann zu schwerwiegenden Konsequenzen für die Eigenschaften der verwendeten statistischen Verfahren wie Erwartungstreue, Signifikanzniveau, Fehlklassifikationswahrscheinlichkeit u. a. führen. Daher ist es oft ratsam, die Gültigkeit

der Verteilungsannahmen anhand der vorliegenden Daten zu überprüfen. Die Autoren sind sich entweder des Unsinns, den sie bedienen, nicht bewusst oder sie nehmen diese Fehlanwendung billigend in Kauf. Es wird nicht untersucht, ob das Modell der Regression angewendet werden kann – es kann nicht angewendet werden. Bei einem Vergleich der Interventionsgruppe (ICM) mit einer Kontrollgruppe (SCM) auf einer Bewertung, die auf der Basis aussagefreier Regressionskoeffizienten beruht, handelt es sich bei den Autoren (sie seien hier genannt: Huber, Reich Früh, Rosemann) um eine dolose Handlung.

9. Signifikanz eines Unterschiedes

Sieht man einmal davon ab, dass der Vergleich der Interventionsgruppe mit der Vergleichsgruppe schon in Bezug auf die „Diagnose" nicht sachgerecht erfolgt, dass insbesondere Verweilzeiten in der Interventionsgruppe schon durch die Auswahl kürzer ausfallen müssen als in der Kontrollgruppe, verwenden die Autoren den Begriff Signifikanz. Dieser Begriff gehört offensichtlich nicht zu dem Bereich, in dem sich die Autoren auskennen. Signifikanz bezieht sich auf die Abweichung im Rahmen eines Modells. Es gibt keine signifikanten Unterschiede *an sich*. Signifikanz ist ohne die Verwendung eines Signifikanzniveaus nicht ausdrückbar. Ein Unterschied, der bei einem Signifikanzniveau α signifikant ist, ist es auch bei jedem größeren Signifikanzniveau.

Ein Unterschied ist signifikant auf dem Signifikanzniveau von z. B. 5 %, wenn die Wahrscheinlichkeit, dass der Unterschied durch Zufall zustande kam, kleiner als 5 % ist – ja und nun?

„Signifikante Ergebnisse" haben eine große Bedeutung in der Veröffentlichungspolitik von Fachzeitschriften in vielen Disziplinen. Diese „Veröffentlichungspolitik verstellt zum Teil den Blick auf andere wichtige Aspekte, zum Beispiel, ob das statistisch signifikante Ergebnis überhaupt irgendeine Frage von Relevanz beantwortet."[2] Innovative Ergebnisse wird man nicht im „International Journal of Integrated Care" erwarten können.

Es ist unakzeptabel einfach, falsche Hypothesen statistisch signifikant als richtig zu belegen. Eine Abhilfe ist auch nur bedingt möglich, wenn die Auswahl der Stichprobe angegeben wird, wenn benannt wird, welche Daten bei der Auswertung ausgelassen wurden, welche Daten nacherhoben wurden und wie, welche Daten nicht zur Verfügung standen aber zur Verfügung hätten stehen müssen, welche Manipulationen angewendet wurden, …

10. p-Wert und Unterschied

Der p-Wert gibt an, wie wahrscheinlich es ist, dass ein „ermittelter" Unterschied zufällig auftritt. Der p-Wert beschreibt nicht irgendeine Fehlerwahrscheinlichkeit, sondern mit welcher Wahrscheinlichkeit die

---

[2] Hans-Peter Beck-Bornholdt, Hans-Hermann Dubben: Der Hund , der Eier legt. Erkennen von Fehlinformationen durch Querdenken, Rowohlt Taschenbuch Verlag, Reinbek 1999

beobachteten Unterschiede auftreten, wenn die miteinander verglichenen Behandlungsprozeduren (bezogen z. B. auf die Verweilzeit) gleich sind. Es kann nicht rückwärts darauf geschlossen werden, dass eine Prozedur (bezogen z. B. auf die Verweilzeit) kürzer ist – auch wenn es scheinbar dem gesunden Menschenverstand widerspricht. Lediglich die Vermutung, dass an dem Unterschied etwas dran sein könnte, wird genährt. Ohne weitere Untersuchungen ist hier keine Bestätigung zu erfahren. Auf diese weiteren Untersuchungen zu verzichten, insbesondere auch die vielen anderen fachlichen Fehler und methodischen Stockfehler nicht zu verringern, ist verantwortungslos.

11. Zusammenfassung

Die Studie als Beweismittel, wonach Managed Care Modelle mit Budgetverantwortung für ihre Patienten kürzere Verweilzeiten bei der Hospitalisierung aufweisen, weil sie besser arbeiten, zu „verkaufen", ist kriminell.

Dass die Auswahl von mehr zu Chronikern erklärten (u. U. gar nicht so sehr kranken) Patienten im Durchschnitt gegenüber den akut erkrankten Patienten und den u. U. Wiederhospitalisierten bzw. später Hospitalisierten im Mittel kostengünstiger gemanagt werden können, ist allgemein bekannt. Durch solche Studien wird weder belegt, dass die Chronifizierung von Krankheiten zunimmt, noch dass selektive Managed Care Modelle mit Budgetverantwortung mehr als nur Mengen- und summarische Kostenerhöhungen bringen.

Zossen, 26.08.2016

Dr. Walter Warmuth

HDoz Dr. rer. nat. habil. Walter Warmuth | Drosselgasse 5 | 15806 Zossen | **T** +49 3377 303621 | **F** +49 3377303622 | **E** ww@suljee.de