

*Interdisziplinäre Forschung  
für Effizienz und Qualität  
im Gesundheitswesen*

**Plädoyer für eine effektivere Nutzung der Datenbestände in  
Deutschland für Wissenschaft und Patienteninformation**

**Expertise im Auftrag der Bertelsmann Stiftung für das Fachgespräch**

**„Datennutzung im Gesundheitswesen“**

Prof. Dr. Jonas Schreyögg , Hamburg Center for Health Economics, Universität Hamburg

## Zusammenfassung

Derzeit wird viel über Big Data im Gesundheitswesen geredet – vieles davon ist Zukunftsmusik. Dabei weist Deutschland schon jetzt eine hervorragende Datenbasis im Gesundheitswesen auf. Prinzipiell stehen eine Reihe von Datensätzen, unter anderem stationäre Abrechnungsdaten und Kassendaten, für wissenschaftliche Zwecke und die Patienteninformation zur Verfügung. Allerdings existieren hinsichtlich der konkreten Freigabe und Nutzung dieser Datensätze große administrative Hürden und Vorbehalte, die überwiegend nicht auf Datenschutzgesetze zurückzuführen sind. Unter anderem würde eine pseudonymisierte Bereitstellung von 70-Prozent-Stichproben für Nutzer dieser Datensätze beim DIMDI oder bei den Forschungsdatenzentren der Länder die Situation deutlich verbessern. Umgesetzt werden könnte dies durch Rechtsverordnungen oder Änderungen in bestehenden Rechtsverordnungen. Andere Datensätze zu Leistungs- und Strukturdaten der Leistungserbringer liegen grundsätzlich vor, sind aber nicht öffentlich zugänglich (ambulant) bzw. sollten erweitert werden (stationär). Hier ist der Gesetzgeber aufgerufen, die Voraussetzungen für eine Nutzung zu schaffen. Insgesamt hätten verbesserte Nutzungsmöglichkeiten der Datenbestände einen erheblichen Mehrwert für Leistungsanbieter, Kostenträger, PatientInnen, Politik und Verbände.

## Ausgangslage

Viele durch das Gesundheitssystem generierte Systemdaten (u. a. Routedaten sowie Struktur- und Leistungsdaten) liegen bei verschiedenen Trägern und Dienstleistern vor. Ihre Qualität ist im Vergleich zu ähnlichen Daten in anderen Ländern als sehr gut einzustufen. Kaum ein Land weist eine so hervorragende Datenbasis wie Deutschland auf. Allerdings können diese Daten hierzulande nur sehr eingeschränkt genutzt werden. Es bestehen große administrative Hürden bei der Nutzung für wissenschaftliche Zwecke und zur Patienteninformation, die überwiegend nicht auf Datenschutzgesetze zurückzuführen sind. Dabei wären diese Daten für die Versorgungsforschung, Versorgungssteuerung und Patienteninformation ein wertvoller Schatz. Bessere Nutzungsmöglichkeiten der Datenbestände könnten einen erheblichen Mehrwert für PatientInnen, Politik und Verbände schaffen.

Ziel dieses Papiers ist, 1) über die derzeitige Situation der Nutzung von Daten für wissenschaftliche Zwecke und zur Patienteninformation zu informieren, 2) den potenziellen gesellschaftlichen Mehrwert von besseren Nutzungsmöglichkeiten darzustellen und 3) Maßnahmen zu empfehlen, um diese Nutzung zu ermöglichen bzw. zu erleichtern.

## Welchen gesellschaftlichen Mehrwert hätte ein angemessener Zugang zu den Daten?

1) **Mehrwert für Politik und Verbände:** In den vergangenen Jahren ist ein klarer Trend zu evidenzbasierter Gesundheitspolitik in Deutschland erkennbar. Gerade bei der Konzeption von Reformen ist es wichtig abzuschätzen, wie diese Reformen wirken werden und ob gegebenenfalls nach einigen Jahren eine Nachsteuerung erforderlich ist. Dabei verweisen Politik und Verbände regelmäßig zu Recht auf die mangelnde wissenschaftlichen Evidenz in vielen Bereichen des Gesundheitswesens. Die Bereitstellung von systematisch aufbereiteten Informationen und darauf aufbauenden Analysen führt oftmals zu einer Versachlichung und einer ausgewogeneren Darstellung der Thematik in der Öffentlichkeit.

Beispiele der letzten Jahre sind hier das Krankengeld und die Fallzahlentwicklung im stationären Sektor. In der aktuellen Debatte um Upcoding im ambulanten Sektor könnte dies ebenfalls helfen. Solide Studien erfordern allerdings neben einer qualitativ guten Datenbasis, die hierzulande in der Regel gegeben ist, einen angemessenen Zugang zu diesen Daten für wissenschaftliche Zwecke.

2) **Mehrwert für PatientInnen:** Verbesserte Nutzungsmöglichkeiten der Daten könnten aus Perspektive der PatientInnen zu mehr Transparenz im Gesundheitswesen beitragen. Daten, die Auskunft über den Nutzen sowie die Qualität von Leistungen und Leistungserbringern im Gesundheitswesen geben können, sind eine wichtige Grundlage, um informierte Entscheidungen über die Art der Behandlung oder den Behandler zu treffen. Dabei geht es nicht nur um Daten aus der Qualitätssicherung, sondern auch um Struktur- und Leistungsdaten (z. B. medizinische Schwerpunkte, qualitätsgesicherte Leistungen), die bei der Auswahl eines Leistungserbringers für Patienten von Bedeutung sein können.

3) **Mehrwert für die Versorgungsforschung:** Die wegweisenden Initiativen der Bundesregierung aus den vergangenen Jahren zur Förderung von Versorgungsforschung und Gesundheitsökonomie (BMBF, BMG, Innovationsfonds) haben in vielen Ländern der Welt große Beachtung gefunden. Ziel ist es, Versorgungsforschung und gesundheitsökonomische Forschung an die Weltspitze heranzuführen. Gerade die weltweit besten Journals im Bereich Versorgungsforschung enthalten allerdings größtenteils Papiere, die auf Routinedaten basieren, etwa das Journal „Health Services Research“. Deutsche VersorgungsforscherInnen und GesundheitsökonomInnen weichen daher zunehmend auf Routinedaten anderer Länder aus, vor allem der skandinavischen Länder, Englands und der USA. Auch wenn dies im Rahmen der Globalisierung unvermeidbar und auch wünschenswert ist, wäre der Nutzen der Ergebnisse hierzulande finanziert Versorgungsforschung für die Bevölkerung größer, wenn Routinedaten aus dem deutschen Gesundheitssystem genutzt werden könnten. Auch würde dies für deutsche VersorgungsforscherInnen die Möglichkeit deutlich verbessern, Mittel aus dem europäischen Forschungsrahmenprogramm „Horizon 2020“ einzuwerben.

### **Um welche Daten geht es?**

Es geht insbesondere um Routinedaten, die zum Zweck der Abrechnung im Gesundheitssystem entstehen, sowie um Struktur- und Leistungsdaten, die ohnehin erhoben werden. Die wichtigsten Datenbestände:

1) **stationäre Abrechnungsdaten:** Daten gemäß § 21 KHEntgG, die vom InEK an das Statistische Bundesamt übermittelt werden

2) **Kassendaten:** u. a. stationäre Abrechnungsdaten gemäß § 301 SGB V, die im Unterschied zu den Daten gemäß § 21 KHEntgG eine vor- und nachstationäre Perspektive ermöglichen; ambulante Abrechnungsdaten gemäß § 295 SGB V; Daten zur Arzneimittelversorgung gemäß § 300 SGB V i.V.m. § 31 SGB V.

3) **Struktur- und Leistungsdaten der ambulanten und stationären Leistungserbringer:** Im ambulanten wie auch im stationären Sektor liegen diese Informationen prinzipiell vor, allerdings sind sie nur teilweise öffentlich zugänglich, etwa über die strukturierten Qualitätsberichte.

### **Welche Daten können jetzt schon genutzt werden und worin bestehen die Hürden bei der Datennutzung?**

In den vergangenen Jahren wurden die stationären Abrechnungsdaten gemäß § 21 KHEntgG sowie die Kassendaten prinzipiell – für wissenschaftliche Zwecke anonymisiert – als Individualdaten (d. h. mit Einzelfallbezug) zur Verfügung gestellt. Die stationären Abrechnungsdaten gemäß § 21 KHEntgG sind seit einigen Jahren als DRG-Statistik im Forschungsdatenzentrum (FDZ) Bund verfügbar. Dies sind die Daten, die das Institut für das Entgeltsystem im Krankenhaus (InEK) sammelt und nutzt. Auch die Informationen aus den Daten gemäß § 295 SGB V, § 300 SGB V i.V.m. § 31 SGB V und § 301 SGB V stehen größtenteils über § 303a–e SGB V beim Deutschen Institut für Medizinische Dokumentation und Information (DIMDI) für Forschungszwecke zur Verfügung. Dies sind die Daten, die das Bundesversicherungsamt (BVA) für den Morbi-RSA nutzt. Im Folgenden werden die beiden Datenbestände vereinfacht als „DRG-Statistik“ und „Kassendaten“ bezeichnet. Auch wenn die Daten formal für wissenschaftliche Zwecke verfügbar sind, verhindert die Form der Bereitstellung aus verschiedenen Gründen eine adäquate wissenschaftliche Nutzung.

1) **Umfang der bereitgestellten Informationen:** In beiden Datenbeständen fehlen maßgebliche Informationen, die für viele Forschungsfragen ebenso erforderlich wären wie für eine verbesserte Leistungs- und Qualitätstransparenz.

Bei den *Kassendaten* fehlen bestimmte Informationen, die nicht von den Kassen an das BVA gemeldet werden. Dies sind Kennziffern für die stationären und ambulanten Leistungserbringer – man kann also keine auf Leistungserbringer bezogene Fragestellung mit diesen Daten bearbeiten. So kann etwa die Frage, ob sich die poststationäre Mortalität/Wiedereinweisung von Patienten nach Regionen, Trägern oder Größenklassen der Krankenhäuser unterscheidet, nicht beantwortet werden. Dies wäre aber aus versorgungspolitischer Perspektive und auch im Sinne der Patienteninformation relevant. Darüber hinaus sind die Daten nicht als Panel zu nutzen – man kann also (anonymisierte) Informationen zu Patienten nicht über mehr als ein Jahr nachverfolgen. Dies ist jedoch für viele Fragestellungen, beispielsweise für Analysen im Kontext chronischer Erkrankungen, sehr wichtig.

In der *DRG-Statistik* ist die Problematik anders gelagert. Es werden im Prinzip alle notwendigen Daten vom InEK an das Statistische Bundesamt übermittelt. Dort jedoch werden diverse Variablen von der Nutzung ausgeschlossen. Darunter fällt etwa die Variable zur Trägerschaft. Das Statistische Bundesamt argumentiert, diese Variablen seien nicht valide. Diese Aussage kann jedoch nicht überzeugen, da beispielsweise die gleiche Trägerschaftsvariable in der Krankenhausstatistik freigegeben ist. Auch darf der

öffentlich verfügbare und offiziell vom InEK zertifizierte Grouper – von den Krankenhäusern zur Eingabe von Fällen und Ermittlung der Fallpauschalen genutzt – nicht zum Datensatz zugespielt werden, was dazu führt, dass die Daten nicht über die Jahre vergleichbar sind. Das bedeutet also, dass nahezu alle das DRG-System betreffenden Fragestellungen so nicht bearbeitet werden können. Eine weitere deutliche Einschränkung ist, dass das Hinzuspielen externer Informationen auf sehr wenige Variablen begrenzt ist und die entsprechenden Datensätze öffentlich sein müssen. Beispielsweise können Ergebnisse von Krankenhausbefragungen zu institutionellen Gegebenheiten nur unter sehr restriktiven Bedingungen zum Datensatz hinzugespielt werden.

**2) Prozedere der Datennutzung:** Ein Grundproblem der Datenbereitstellung beider Bestände ist, dass Nutzer nicht auf dem Volldatensatz arbeiten dürfen. Sie müssen die Daten per kontrollierter Ferndatenverarbeitung nutzen. Dies sei an einem Beispiel der DRG-Statistik verdeutlicht:

Der Nutzer kann den anonymisierten Volldatensatz nicht sehen. Er erhält stattdessen einen Beispieldatensatz und entwickelt auf dieser Basis einen Programmiercode einer Statistiksoftware (z. B. SAS), der bei komplexen Projekten 100 Seiten und mehr umfassen kann. Diesen Programmiercode sendet er dann an das FDZ, das ihn auf dem Volldatensatz laufen lässt. Anschließend erhält der Nutzer – nach eigener Erfahrung vier bis acht Wochen später – das Ergebnis. Falls sich auch nur ein kleiner Fehler in dem Programmiercode befindet, bricht das Programm ab und man erhält lediglich eine Fehlermeldung. Das heißt: Man korrigiert den Programmiercode, reißt sich erneut in die Warteschlange ein und muss wiederum mehrere Wochen auf das Ergebnis warten. Ein Forschungsprojekt zu einer komplexen Fragestellung, heutzutage die Regel, dauert auf diese Weise Jahre und ist somit nicht zu realisieren.

Die Idee eines Zugangs über kontrollierte Ferndatenverarbeitung stammt aus den 1990er Jahren. Damals war die Datenverarbeitungskapazität ohnehin sehr gering und auch die Komplexität und Präzision der in der Wissenschaft genutzten Modelle waren noch bei weitem nicht so fortgeschritten wie heute. Um praxisrelevante und wissenschaftlich anschlussfähige Ergebnisse zu produzieren, reicht es heutzutage in der Regel nicht, das arithmetische Mittel für Studien zu berechnen. Vielmehr werden, selbst in gesetzlich definierten Forschungsaufträgen, überwiegend multivariate Analysen möglichst mit kausalem Anspruch gefordert. Dies ist mit der derzeitigen Form des Datenzugangs über das FDZ sowie beim DIMDI nicht möglich.

Sowohl das FDZ als auch das DIMDI begründen die restriktiven Bedingungen der Datennutzung mit Datenschutzbestimmungen. Dieses Argument verwundert jedoch, da es zahlreiche Beispiele der offiziellen Bereitstellung von Daten gibt, die ein direktes Arbeiten auf dem anonymisierten Volldatensatz ermöglichen. So können WissenschaftlerInnen bei einzelnen Krankenkassen auch direkt vor Ort oder per VPN-Tunnel auf dem Datensatz dieselben Daten mit sogar mehr Variablen als den §303 ff. bearbeiten. Auch die Arbeitsmarktstatistik und die Krankenhausstatistik in den Forschungsdatenzentren der Länder können direkt auf dem Volldatensatz und/oder einer 70-Prozent-Stichprobe bearbeitet werden.

3) **Ausleiten von Ergebnissen:** Mit der Ausleitung der Ergebnisse, also dem Zurverfügungstellen der Ergebnisse für die Nutzer, werden die Daten formal öffentlich. An dieser Stelle muss die rein wissenschaftliche Nutzung von der Nutzung zur Patienteninformation unterschieden werden.

Für eine rein wissenschaftliche Nutzung stellen weder der Datenschutz noch die Wahrung des Schutzes von Einrichtungen (Betriebsgeheimnisse) eine Hürde dar. Beispielsweise ist eine Ausleitung von Informationen zu einzelnen Krankenhäusern nicht möglich. Da aus wissenschaftlicher Sicht übergreifende und verallgemeinerbare Aussagen getätigt werden sollen, ist dies kein Problem – für die wissenschaftliche Nutzung werden ohnehin nur aggregierte Darstellungen mit Regressionskoeffizienten ausgeleitet, die keinerlei Rückschluss auf einzelne Individuen oder Institutionen ermöglichen.

Anders verhält es sich bei der Nutzung der Daten zum Zweck der Patienteninformation. Hier wäre eine Darstellung auf Einrichtungsebene, wie sie unter anderem in Großbritannien und den skandinavischen Ländern erlaubt ist, aufgrund des Schutzes von Einrichtungsdaten derzeit nicht möglich. Es müssen immer mindestens zehn Einrichtungen zusammengefasst werden. Auch hier stellt sich die Problematik des Datenschutzes nicht, da es um einrichtungsbezogene Informationen geht. Diese weisen ohnehin nur einen Informationswert auf, wenn es um eine substantielle Fallzahl pro Einrichtung geht.

Die *drei genannten Problemfelder zusammengenommen*, bedeuten, dass die qualitativ an sich guten Daten aufgrund von Defiziten bei der Bereitstellung aus wissenschaftlicher Sicht deutliche Mängel aufweisen und nur für sehr wenige wissenschaftliche Fragestellungen genutzt werden können. Auf diesen Missstand machen regelmäßig verschiedene wissenschaftliche Gesellschaften aufmerksam (u. a. Deutsches Netzwerk für Versorgungsforschung, Deutsche Gesellschaft für Gesundheitsökonomie). Auch der Sachverständigenrat (SVR) hat darauf verschiedentlich hingewiesen (SVR 2012; SVR 2015). Die Nutzerzahlen der Statistiken belegen, dass die Nutzung der entsprechenden Daten unter diesen Bedingungen nicht attraktiv ist. Das DIMDI berichtet von nur 38 Anträgen zwischen 2014 und 2016 (DIMDI 2016). In dem Bericht wird konstatiert, dass die Gründe vor allem in der unzureichenden Nutzbarkeit und dem mangelnden Nutzen des Angebots liegen. Gegen die Nutzung für die Patienteninformation wird der Schutz der Einrichtungsdaten ins Feld geführt.

Bei den Struktur- und Leistungsdaten sieht die Problematik etwas anders aus. Hier liegen zu den stationären Leistungserbringern bereits einige Informationen in den strukturierten Qualitätsberichten öffentlich vor. Diese sind allerdings aufgrund der heterogenen Datenqualität nur nach Aufbereitung und Plausibilisierung aus Patientenperspektive nutzbar. Im ambulanten Bereich liegen diese Informationen grundsätzlich vor, sind aber nicht öffentlich zugänglich.

### **Wie ist die Situation in anderen Ländern?**

In den meisten europäischen Ländern, etwa in England, den Niederlanden oder der Schweiz, dürfen WissenschaftlerInnen direkt auf dem pseudonymisierten Datensatz arbeiten. Dies ist in verschiedenen Projekten, gefördert durch die europäischen Forschungsrahmenprogramme, dokumentiert. Es gibt

auch einen ausführlichen Bericht der Organisation für wirtschaftliche Zusammenarbeit und Entwicklung (OECD), der die Praxis anderer OECD-Länder bei der Nutzung von Daten des Gesundheitswesens beschreibt (OECD 2015). Deutschland ist in dem Bericht nicht erfasst. Es wird deutlich, dass andere Länder trotz ähnlicher Datenschutzregelungen gerade bei der Nutzung und bei den Möglichkeiten der Zuspiegelung anderer Daten liberalere Vorschriften haben oder die Vorschriften vor allem hinsichtlich des *Einrichtungsschutzes* liberaler auslegen.

Bei der Nutzung der Daten für die Patienteninformation existieren große Unterschiede, doch immer mehr Länder öffnen ihre Leistungs- und Strukturdaten sowie ihre Routinedaten für diesen Zweck. Vor allem in England und den skandinavischen Ländern können vielfach administrative Datensätze für die Patienteninformation, auch auf Einrichtungsebene, herangezogen werden.

### **Was wäre wünschenswert und welche Maßnahmen sollten ergriffen werden?**

Bei der Nutzung der Daten für wissenschaftliche Zwecke wie auch für die Patienteninformation wären folgende Maßnahmen, ausgehend vom Status quo, wünschenswert:

1) **DRG-Statistik:** Wichtig wäre die **volle Bereitstellung aller Informationen** aus der Vereinbarung des Selbstverwaltungspartners gemäß § 21 KHEntgG **ohne Ausschluss und Nachbearbeitung bestimmter Variablen** durch das Statistische Bundesamt sowie die **Ermöglichung des Zuspiegels von Software und externen Informationen**. Des Weiteren sollte, analog zu anderen über die Forschungsdatenzentren bereitgestellten Datensätzen, eine 70-Prozent-Stichprobe zur direkten Bearbeitung in den Forschungsdatenzentren der Länder für Forschungszwecke zur Verfügung gestellt werden. Bei der Nutzung der Daten für die Patienteninformation wäre hingegen die Möglichkeit zur Veröffentlichung auf Einrichtungsebene, also der Volldatensatz unabdingbar. Um die Maßnahmen umzusetzen, erscheint keine Änderung gesetzlicher Vorschriften erforderlich. Allerdings wäre eine Anweisung des Bundesministeriums des Innern (Aufsicht) an das Statistische Bundesamt oder eine Rechtsverordnung des Bundesministeriums für Gesundheit notwendig.

2) **Kassendaten:** Bei den Daten gemäß § 303a ff. SGB V wäre zur **Erweiterung des Datensatzes** eine Anpassung der Datenlieferungen von den Krankenkassen an das BVA erforderlich. Hier könnte das Bundesministerium für Gesundheit in § 3 DaTraV (Datentransparenzverordnung) den vom BVA an das DIMDI bereitgestellten Datensatz unter anderem um Institutionenkennziffern sowie zur Nutzung als Panel erweitern. Eine Nutzung als Panel würde eine **Pseudonymisierung** statt einer Anonymisierung des Datensatzes erfordern, verbunden mit der Einrichtung einer **Datenvertrauensstelle**. Dies ist bei der Datennutzung auf Einrichtungsebene im Gesundheitswesen gängige Praxis. Auch hier sollte mindestens eine 70-Prozent-Stichprobe zur Bearbeitung beim DIMDI für Forschungszwecke zur Verfügung gestellt werden. Eine entsprechende Änderung in § 5 DaTraV durch das Bundesministerium für Gesundheit wäre dafür erforderlich. Bei der Nutzung der Daten für die Patienteninformation wäre hingegen die Möglichkeit zur Veröffentlichung auf Einrichtungsebene und somit der Volldatensatz unabdingbar.

3) **Struktur- und Leistungsdaten der ambulanten und stationären Leistungserbringer:** Die für den stationären Bereich öffentlich vorliegenden Informationen aus den strukturierten Qualitätsberichten sollten im Bereich der externen Qualitätssicherung um Routedaten ergänzt werden, wie es der SVR (z. B. SVR 2012) regelmäßig empfiehlt. Für den ambulanten Bereich sollten die Ärzte analoge Informationen über die Leistungserbringer – u. a. **Fachgebiet, Schwerpunkte nach Weiterbildungsordnung, qualitätsgesicherte bzw. genehmigungspflichtige Leistungen** – ebenfalls für die Forschung und die Patienteninformation verfügbar machen. Ein jährlich aktualisierter Datensatz hierzu sollte öffentlich zugänglich gemacht werden. Dafür wäre es erforderlich, dass eine gesetzliche Grundlage geschaffen wird.

Insgesamt wäre ein *angemessener Zugang* zu diesen und anderen Daten für wissenschaftliche Zwecke und für die Patienteninformation wichtig. Nur durch den verbesserten Zugang zu diesen Datenbeständen kann sich hierzulande ein wissenschaftlicher Wettbewerb wie in anderen Ländern, etwa den USA oder England, entwickeln. Von diesem wissenschaftlichen Wettbewerb und seinen Ergebnissen profitieren mittel- bis langfristig alle Akteure des Gesundheitswesens und vor allem die PatientInnen. Letztere würden darüber hinaus auf direktem Wege von der Bereitstellung zusätzlicher Informationen auf Einrichtungsebene profitieren. Die meisten europäischen Länder haben diese Zusammenhänge erkannt und die entsprechenden Daten konsequent für wissenschaftliche Zwecke zur Verfügung gestellt. Im Sinne einer optimalen Verwendung von Ressourcen im Gesundheitswesen sollte Deutschland ihrem Beispiel folgen.

## Literatur

OECD – Organisation for Economic Co-operation and Development (2015): Health Data Governance, Health Policy Studies. Paris.

DIMDI – Deutsches Institut für Medizinische Dokumentation und Information (2016): Informationssystem Versorgungsdaten, Evaluationsbericht des DIMDI gemäß DaTraV § 8.

SVR – Sachverständigenrat zur Begutachtung der Entwicklung im Gesundheitswesen (2012): Wettbewerb an der Schnittstelle zwischen ambulanter und stationärer Gesundheitsversorgung. Mannheim.

SVR – Sachverständigenrat zur Begutachtung der Entwicklung im Gesundheitswesen (2015): Krankengeld – Entwicklung, Ursachen und Steuerungsmöglichkeiten. Göttingen.