

Post-Pandemic Populations

Datenerfordernisse und -defizite in der gegenwärtigen Forschungsdateninfrastruktur

Andreas Backhaus

Bundesinstitut für Bevölkerungsforschung

Veröffentlicht am

27. Mai 2021

Für weitere Informationen

Population Europe

WissenschaftsForum
Markgrafenstraße 37
10117 Berlin, Germany

Mediananfragen

+49 (0)30 2061 383 - 30
office@population-europe.eu

Datenerfordernisse und -defizite in der gegenwärtigen Forschungsdateninfrastruktur

Andreas Backhaus

Bundesinstitut für Bevölkerungsforschung

Empfehlungen

- **Die Erfassung der Sterbefälle und der Todesursachen sollte verbessert werden.**
- **Die Konsistenz des Datenangebots sollte sichergestellt werden.**
- **Synergien zwischen vorhandenen Forschungsdaten sollten stärker genutzt werden.**
- **Die Forschungsdateninfrastruktur sollte mit den gewachsenen Möglichkeiten und Erfordernissen evidenzbasierter Forschung Schritt halten.**

Die Forschungsdateninfrastruktur Deutschlands und ihre Defizite sind seit mehreren Jahren ein bedeutendes Thema in der wissenschaftlichen Gemeinschaft. So betonten die Nationale Akademie der Wissenschaften Leopoldina et al. (2015) bereits die Bedeutung bevölkerungsweiter Längsschnittstudien für Wissenschaft und Gesellschaft. Exemplarisch für mehrere wissenschaftliche Vereinigungen setzte sich der Verein für Socialpolitik im darauffolgenden Jahr erfolgreich für eine längere Speicherfrist der Unternehmensindikatoren in administrativen Datensätzen ein (VfS, 2016). Der Rat für Sozial- und Wirtschaftsdaten (RatSWD) sprach 2017 Empfehlungen für die sozial-, verhaltens- und wirtschaftswissenschaftliche Survey-Landschaft in Deutschland aus (RatSWD, 2017).

Durch die COVID-19-Pandemie hat die Thematik der Forschungsdateninfrastruktur einen neuen und zusätzlichen Auftrieb erhalten. Erstens hat die Pandemie ein breiteres Bewusstsein für mehrere, bereits seit längerem bekannte Defizite geschaffen. Dies betrifft vor allem die zeitnahe und standardisierte Erfassung der Mortalität in Deutschland. Zweitens stellt die Pandemie neuartige Herausforderungen an bestehende Strukturen und Abläufe der Sammlung und Bereitstellung von Forschungsdaten. Drittens müssen

Zugänglichkeit und Tiefe des Datenangebots mit der sich weiterentwickelnden Methodik der empirischen Sozialwissenschaften Schritt halten.

Eine zeitnahe und standardisierte Mortalitätserfassung

Die zeitnahe, umfassende und präzise Erfassung der Mortalität ist sowohl für die zukünftige Forschung zu demografischen Folgen der COVID-19-Pandemie als auch für die gesellschaftliche Debatte, welche die Pandemie stets begleitet hat, von Relevanz. So führen Verzögerungen bei der Erfassung von Sterbefällen im öffentlichen Diskurs im Rahmen der COVID-19-Pandemie immer wieder zu verfrühten Meldungen über vermeintliche sinkende Todeszahlen (Backhaus, 2020).

Das Problem der Mortalitätserfassung und entsprechende Lösungsvorschläge werden im Rahmen der COVID-19-Pandemie bereits breit auf internationaler Ebene diskutiert (Setel et al., 2020; Vital Strategies und World Health Organization, 2020), begleitet von Empfehlungen für die Implementierung so genannter „Rapid Mortality Surveillance“ (RMS). Mit Blick auf

Deutschland identifizieren Schneider et al. (2020) drei Defizite der gegenwärtigen Sterblichkeitserfassung:

1. Die zu langen Informationsflüsse von lokalen Gesundheitsämtern und Meldeeinrichtungen zum Robert Koch-Institut, den Statistischen Ämtern des Bundes und der Länder oder anderen zentralen Einrichtungen.
2. Die fehlende zentrale Erfassung des allgemeinen Sterblichkeitsgeschehens und der Todesursachen.
3. Den verbesserungswürdigen Zugang von Forschung und Öffentlichkeit zu aussagekräftigen Mortalitätsdaten. Schneider et al. (2020) empfehlen den vollständigen Wechsel auf elektronische Systeme zur Mortalitätsmeldung, den Aufbau eines Nationalen Mortalitätsregisters sowie die Nutzung eines stochastischen Überlagerungsverfahrens, um auch kleine Fallzahlen gesichert anonymisiert publizieren zu können.

Ein konsistentes und konsolidiertes Datenangebot

Der Ausbruch der COVID-19-Pandemie hat im historischen Vergleich zu einer bisher nicht gekannten Dokumentation der Pandemie und ihrer Folgen durch neue und schnell bereitgestellte Daten geführt. Auch demografische und sozialwissenschaftliche Forschungseinrichtungen haben begonnen, entsprechende Daten und auf diese gestützte Evidenz zu liefern. So haben mehrere bereits regelmäßig in Deutschland erhobene Surveys Sonderbefragungen mit speziellem Bezug zu COVID-19 eingerichtet und/oder ihren Fragenkatalog um entsprechende Module erweitert. Allerdings haben die europaweiten Kontaktbeschränkungen die Erhebung neuer Daten erheblich erschwert, da die Möglichkeiten zur persönlichen Befragung der Teilnehmer*innen überwiegend nicht mehr gegeben waren. Als Konsequenz sind Surveys verstärkt zur telefon- und computergestützten Befragung übergegangen. Dokumentiert sind diese Schritte u. a. für den „Survey of Health, Ageing and Retirement in Europe“ (SHARE) (Scherpenzeel et al., 2020) und für das „German Family Panel“ (pairfam) und den „Generations and Gender Survey“ im

Rahmen der FReDA-Infrastruktur (Gummer et al., 2020). Stellenweise wurden existierende Survey-Strukturen auch genutzt, um gänzlich neue, rein auf die COVID-19-Pandemie ausgerichtete Surveys zu starten, so im Fall der Mannheimer Corona-Studie, welche auf dem „German Internet Panel“ aufbaut (Blom et al., 2020).

Eine Herausforderung, die sich aus diesem neuen Datenangebot ergibt, ist die Sicherstellung der Vergleichbarkeit und ggf. auch der Verknüpfbarkeit der neu erhobenen Daten mit den in der Forschungsdateninfrastruktur existierenden Beständen. Daten, die während der COVID-19-Pandemie erhoben werden, sollten idealerweise mit bereits existierenden Daten, die die Situation vor der Pandemie abbilden, in Verbindung gebracht werden können, um vorher-nachher-Vergleiche zu ermöglichen. Dieser Anspruch kann gelockert werden, wenn Aggregate und nicht Individuen im Mittelpunkt der Analyse stehen; in diesem Fall muss lediglich die Repräsentativität der Datenquellen vor und während der Pandemie sichergestellt sein. So greifen beispielsweise Bujard et al. (2020) auf Daten des Mikrozensus aus den Vorjahren zurück, um Beschäftigungs- und Familiensituation verschiedener Bevölkerungsgruppen vor der Pandemie zu charakterisieren und diese dann mit den während der Pandemie ermittelten Werten zu vergleichen.

Während die Verknüpfbarkeit der Daten auf aggregierter Ebene somit ein geringes Problem darstellt, werden auf Individualebene nicht immer diejenigen Personen befragt, die bereits in früheren Survey-Runden zur Verfügung standen, oder es wurde eine neue Befragung anlässlich der Pandemie begonnen. In diesen Situationen stehen generell keine Individualdaten aus der Zeit vor der Pandemie zur Verfügung. Einige Surveys behelfen sich deshalb mit retrospektiven Befragungen zur Situation vor Ausbruch der Pandemie. Die dabei erhobenen Informationen spiegeln jedoch teilweise subjektive Einschätzungen wider und sind nicht so präzise wie beispielsweise administrative Daten. Eine Lösung könnte hier ein Ausbau der Verlinkungen innerhalb der Forschungsdateninfrastruktur darstellen, d. h. die Verknüpfung von Individualdaten einzelner Surveys mit größeren, administrativen Datenbanken. Eine solche Verlinkung besteht beispielsweise schon zwischen in Deutschland erhobenen SHARE-Daten und den Daten der Deutschen Rentenversicherung (Börsch-Supan et al., 2020).

Egger (2020) präsentiert eine aktuelle Zusammenstellung von in Deutschland erhobenen Mikrodatsätzen mit vorwiegend wirtschaftswissenschaftlichem Bezug. Unter diesen bietet das Forschungsdatenzentrum des Instituts für Arbeitsmarkt- und Berufsforschung (IAB) einen sowohl im deutschen als auch im internationalen Vergleich sehr umfangreichen Datensatz zu sozialversicherungspflichtigen Beschäftigten an (Müller und Wolter, 2020), welcher sich für weitere Verlinkungen anbieten könnte.

An dieser Stelle sollen zwei Beispiele aus der aktuellen Forschung die Potenziale einer umfangreichen, zugänglichen und gleichzeitig sicheren Forschungsdateninfrastruktur unterstreichen: Alstadsæter, Bjørkheim et al. (2020) nutzen norwegische Registerdaten, um annähernd in Echtzeit die durch die Pandemie in Norwegen verursachten Entlassungen zu dokumentieren, wobei sie letztere nach Umfang, Dynamik und sozioökonomischen Charakteristika der Entlassenen differenzieren können. In diesen Registerdaten sind Informationen über 2,67 Millionen der insgesamt 5,37 Millionen Einwohner*innen Norwegens enthalten. Ebenfalls auf Basis umfangreicher norwegischer Daten schätzen Alstadsæter, Bratsberg et al. (2020) Stärke und Verteilung der Umsatzeinbrüche der dortigen Unternehmen während der Pandemie und können zudem die Wirkungen verschiedener Maßnahmen zur Unternehmenshilfe simulieren. Für diese Analyse konnten Daten von etwa 158.000 norwegischen Unternehmen herangezogen werden.

Eine Forschungsdateninfrastruktur für evidenzbasierte Forschung

Die im Zuge der COVID-19-Pandemie ergriffenen politischen Maßnahmen haben weitreichende Auswirkungen auf das soziale und wirtschaftliche Leben der jeweiligen Gesellschaften. Während beispielsweise Kontaktbeschränkungen, Reisewarnungen und im US-amerikanischen Raum sogenannte „stay-at-home“-Anweisungen darauf abzielen, das soziale und damit auch zwingenderweise das wirtschaftliche Leben herunterzufahren, um das Infektionsgeschehen zu verlangsamen und zu unterbrechen, treten begleitend oft konjunktur- und sozialpolitische Maßnahmen in Kraft, um die wirtschaftlichen Folgen der gesundheitspolitischen Eingriffe abzumildern.

Die Wirksamkeit sowohl der gesundheits- als auch der wirtschaftspolitischen Maßnahmen steht aktuell im Mittelpunkt der evidenzbasierten Forschung. Letzterer stellt ein in den letzten Jahren stark gewachsenes Instrumentarium an empirischen Methoden zur Verfügung, um kausale Wirkungszusammenhänge zwischen Ereignissen verschiedenster Art und den jeweils zu beeinflussenden Ergebnissen zu schätzen. Die Herausforderung bei diesen Schätzungen besteht insbesondere darin, dass die betreffenden politischen Interventionen typischerweise nicht nach dem Charakter einer randomisierten kontrollierten Studie erfolgen, sondern eher dort und dann ergriffen werden, wo es besonders notwendig erscheint. Dies hat u. a. zur Folge, dass die Beobachtungseinheiten, die zu einem gegebenen Zeitpunkt einer politischen Maßnahme oder einem anders gearteten „Treatment“ unterliegen, grundsätzlich von vorneherein nicht mit denjenigen Beobachtungseinheiten vergleichbar sind, die dieser Maßnahme nicht unterliegen („Control“). Ein einfacher vorher-nachher-Vergleich der Ergebnisse beider Gruppen ist daher nicht geeignet, um die kausale Wirkung der Maßnahme herauszustellen. Instrumentalvariablen, der Differenzen-in-Differenzen-Ansatz sowie sogenannte „Event Studies“ sind einige der in den empirischen Sozialwissenschaften verbreiteten Methoden, die es unter variierenden Annahmen erlauben, kausale Wirkungszusammenhänge auch in nicht-experimentellen Datensätzen zu identifizieren. Einen einleitenden methodischen Überblick bieten Angrist und Pischke (2008, 2014) im Allgemeinen sowie Goodman-Bacon und Marcus (2020) mit direktem Bezug zur COVID-19-Pandemie. Auch demografische Auswirkungen der Pandemie werden mit diesen und ähnlichen Methoden empirisch untersucht. Für die deutschsprachigen Wirtschaftswissenschaften hat der Verein für Socialpolitik bereits entsprechende Leitlinien und Empfehlungen veröffentlicht (VfS, 2015).

Teilweise hängen die Datenerfordernisse für derartige empirische Methoden unmittelbar mit den bereits genannten Aspekten zusammen und stellen daher ähnliche Ansprüche an die Forschungsdateninfrastruktur. Beispielsweise wird die Validität des Differenzen-in-Differenzen-Ansatzes hinsichtlich einer bestimmten Forschungsfrage üblicherweise mit der Darstellung von Trendlinien der Kontroll- und Treatment-Gruppen vor Beginn des Treatment überprüft. Auch hier sind also konsistente Daten der Studienteil-

nehmer*innen, seien es Individuen, Unternehmen, Familien oder administrative Einheiten, vor und nach Beginn der Intervention erforderlich. Oftmals wird zur Schätzung kausaler Zusammenhänge auch die geografisch und zeitlich unterschiedliche Implementierung von Politikmaßnahmen ausgenutzt, um die Studienpopulation entlang dieser Dimensionen in Kontroll- und Treatment-Gruppen einzuteilen. Hier steht die Forschungsdateninfrastruktur vor der Herausforderung, überhaupt genügend Befragungen und Erhebungen anbieten zu können, die dafür konzipiert worden sind, auch unterhalb der Bundesebene repräsentativ für die entsprechenden Populationen, wie beispielsweise die Bevölkerungen der einzelnen Bundesländer, zu sein. Hier ist die dezentrale, an vielen verschiedenen Instituten und Universitäten angesiedelte Forschungsdatenstruktur als ein Hinderungsgrund zu vermuten. Stark disaggregierte geografische Merkmale wie der Wohnort oder Kreis, in dem eine Person oder ein Unternehmen ansässig sind, sind zudem aus datenschutzrechtlichen Gründen meist auch nicht in den sogenannten „Scientific Use Files“ enthalten, sondern können nur durch die sogenannte „On-Site-Nutzung“ in den entsprechenden Einrichtungen abgefragt werden. Einen aktuellen Überblick zum sogenannten „Remote Access“ zu Daten der amtlichen Statistik und der Sozialversicherungsträger liefert RatSWD (2019). In dieser Hinsicht impliziert die dezentrale Ansiedlung der Forschungsdateninfrastruktur ggf. einen nicht zu unterschätzenden Reiseaufwand für Forscher*innen, der durch die COVID-19-Pandemie zusätzlich erhöht worden ist.

Literatur

- Alstadsæter, A., Bjørkheim, J. B., Kopczuk, W. & Økland, A. (2020). *Norwegian and US policies alleviate business vulnerability due to the Covid-19 shock equally well* (NBER Working Paper 27367). <http://www.nber.org/papers/w27367>
- Alstadsæter, A., Bratsberg, B., Eielsen, G., Kopczuk, W., Markussen, S., Raaum, O. & Røed, K. (2020). *The first weeks of the coronavirus crisis: Who got hit, when and why? Evidence from Norway* (NBER Working Paper 27131). <https://www.nber.org/papers/w27131>
- Angrist, J. D. & Pischke, J. S. (2008). *Mostly harmless econometrics: An empiricist's companion*. Princeton University Press.
- Angrist, J. D. & Pischke, J. S. (2014). *Mastering 'Metrics: The Path from Cause to Effect*. Princeton University Press.
- Backhaus, A. (2020). Common Pitfalls in the Interpretation of COVID-19 Data and Statistics. *Intereconomics*, 55, 162–166. <https://doi.org/10.1007/s10272-020-0893-1>
- Blom, A. G., Wenz, A., Rettig, T., Reifenscheid, M., Naumann, E., Möhring, K., ... & Fikel, M. (2020). *Die Mannheimer Corona-Studie: Das Leben in Deutschland im Ausnahmezustand: Bericht zur Lage vom 20. März bis 30. Juni 2020*.
- Börsch-Supan, A., Czapliski, C., Friedel, S., Herold, I., Korbmayer, J. & Mika, T. (2020). SHARE-RV: Linked Data to Study Aging in Germany. *Jahrbücher für Nationalökonomie und Statistik*, 240(1), 121–132. <https://doi.org/10.1515/jbnst-2018-0034>
- Bujard, M., Laß, I., Diabaté, S., Sulak, H. & Schneider, N. F. (2020). *Eltern während der Corona-Krise. Zur Improvisation gezwungen*. Bundesinstitut für Bevölkerungsforschung. Wiesbaden.
- Egger, P. H. (2020). Introduction to the special issue on German micro datasets. *German Economic Review*, 21(3), 271–271. <https://doi.org/10.1515/ger-2020-0053>
- Goodman-Bacon, A. & Marcus, J. (2020). Using Difference-in-Differences to Identify Causal Effects of COVID-19 Policies. *Survey Research Methods*, 14(2), 153–158. <https://doi.org/10.18148/srm/2020.v14i2.7723>
- Gummer, T., Schmiedeberg, C., Bujard, M., Christmann, P., Hank, K., Kunz, T., ..., Neyer, F. J. (2020). The impact of Covid-19 on fieldwork efforts and planning in pairfam and FReDA-GGS. *Survey Research Methods*, 14(2), 223–227. <https://doi.org/10.18148/srm/2020.v14i2.7740>
- Müller, D. & Wolter, S. (2020). German labour market data – Data provision and access for the international scientific community. *German Economic Review*, 21(3), 313–333. <https://doi.org/10.1515/ger-2019-0127>
- Nationale Akademie der Wissenschaften Leopoldina, acatech – Deutsche Akademie der Technikwissenschaften, Union der deutschen Akademien der Wissenschaften (2015). *Wissenschaftliche und gesellschaftspolitische Bedeutung bevölkerungsweiter Längsschnittstudien*. Halle (Saale).

- OECD (2020). *The Territorial Impact of COVID-19: Managing the Crisis Across Levels of Government. OECD Policy Responses to Coronavirus (COVID-19)*. <http://www.oecd.org/coronavirus/policy-responses/the-territorial-impact-of-covid-19-managing-the-crisis-across-levels-of-government-d3e314e1>

- RatSWD [Rat für Sozial- und Wirtschaftsdaten] (2017). *Die sozial-, verhaltens- und wirtschaftswissenschaftliche Survey-Landschaft in Deutschland – Empfehlungen des RatSWD*. RatSWD Output 6(5). Berlin, Rat für Sozial- und Wirtschaftsdaten(RatSWD). <https://doi.org/10.17620/02671.5>.

- RatSWD [Rat für Sozial- und Wirtschaftsdaten] (2019). *Remote Access zu Daten der amtlichen Statistik und der Sozialversicherungsträger*. RatSWD Output 5(6). Berlin, Rat für Sozial- und Wirtschaftsdaten (RatSWD). <https://doi.org/10.17620/02671.42>

- Scherpenzeel, A., Axt, K., Bergmann, M., Douhou, S., Oepen, A., Sand, G., ... & Börsch-Supan, A. (2020). Collecting survey data among the 50+ population during the COVID-19 outbreak: The Survey of Health, Ageing and Retirement in Europe (SHARE). *Survey Research Methods*, 14(2), 217–221. <https://doi.org/10.18148/srm/2020.v14i2.7738>

- Schneider, N. F., Mueller, U. & Klüsener, S. (2020). *Sterblichkeit in Deutschland muss schneller und besser erfasst werden – Lehren aus der Covid-19-Pandemie* (BiB Working Paper 1/2020). Bundesinstitut für Bevölkerungsforschung.

- Setel, P., AbouZahr, C., Atuheire, E. B., Bratschi, M., Cercone, E., Chinganya, O., ... & Karpati, A. (2020). Mortality surveillance during the COVID-19 pandemic. *Bulletin of the World Health Organization*, 98(6), 374. <http://dx.doi.org/10.2471/BLT.20.263194>

- VfS [Verein für Socialpolitik] (2015). Leitlinien und Empfehlungen des Vereins für Socialpolitik für Ex post-Wirkungsanalysen.

- VfS [Verein für Socialpolitik] (2016). Offener Brief zur Novelle des Bundesstatistikgesetzes. Frankfurt.

- Vital Strategies & World Health Organization (2020). *Revealing the Toll of COVID-19: A Technical Package for Rapid Mortality Surveillance and Epidemic Response*. New York: Vital Strategies.

