Health research requires the linking of healthcare-related data

Gesundheitsforschung braucht die Verknüpfung versorgungsnaher Daten

Authors

Wolfgang Hoffmann[1, 2, 3], Neeltje van den Berg[1, 4], Jochen Schmitt[2, 5], Peter Ihle[6], Gabriele Anton[7], Thomas Datzmann[5], Bianca Franke[8], Max Geraedts[9], Simone Gloystein[1], Sabine Hanß[10], Philipp Heinrich[11], Elisa Henke[12], Anne Hennings[8], Peter Heuschmann[13, 14], Thomas Illig[15], Monika Kraus[16], Caroline Lang[5], Gabriele Müller[5], Anne Neumann[5], Steffen Ortmann[17, 18], Vishnu Priya[12], Jens-Peter Reese[19], Ulrich Sax[10], Simone Scheithauer[20], Olaf Schoffer[5], Martin Sedlmayr[12], Sebastian Claudius Semler[21], Enno Swart[22], Frank Wissing[23], Sylke Ruth Zeißig[8, 13, 14], Ralf Heyder[24]

Affiliations

1 Institut für Community Medicine, Universitätsmedizin Greifswald, Greifswald, Germany
2 DNVF, Deutsches Netzwerk Versorgungsforschung e.V., Berlin, Germany
3 DZNE, Deutsches Zentrum für Neurodegenerative Erkrankungen e.V., Standort Rostock/Greifswald, Germany
4 Deutsches Zentrum für Kinder- und Jugendgesundheit (DZKJ), Standort Greifswald/Rostock, Germany
5 Zentrum für Evidenzbasierte Gesundheitsversorgung, Medizinische Fakultät Carl Gustav Carus der Technischen Universität Dresden, Dresden, Germany
6 PMV forschungsgruppe, University of Cologne, Köln, Germany
7 Universität Bielefeld, Medizinische Fakultät Ostwestfalen-Lippe (OWL)/Biobank, Bielefeld, Germany
8 ADT, Arbeitsgemeinschaft Deutscher Tumorzentren e.V., Berlin, Germany
9 Institute for Health Services Research and Clinical Epidemiology, Philipps-Universität Marburg, Marburg, Germany
10 Universitätsmedizin Göttingen, Institut für Medizinische Informatik, Göttingen, Germany
11 Medizinische Fakultät der TU Dresden, Unabhängige Treuhandstelle, Dresden, Germany
12 Medizinische Fakultät und Universitätsklinikum Carl Gustav Carus Dresden, TU Dresden, Institut für Medizinische Informatik und Biometrie, Dresden, Germany
13 Julius-Maximilians-Universität Würzburg, Institut für Klinische Epidemiologie und Biometrie, Würzburg, Germany
14 Universitätsklinikum Würzburg, Institut für medizinische Datenwissenschaften, Würzburg, Germany
15 Medizinische Hochschule Hannover, Hannover Unified Biobank, Hannover, Germany
16 Helmholtz Zentrum München, Institut für Epidemiologie, München, Germany
17 Referat Strukturwandel, Medizinische Fakultät Carl Thiem, Cottbus, Germany
18 München Klinik gGmbH, München Klinik gGmbH, München, Germany
19 Technische Hochschule Mittelhessen (THM), Fachbereich Gesundheit, Gießen, Germany
20 Department of Infection Control and Infectious Diseases, University Medical Center, Georg-August-University, Goettingen, Germany
21 Geschäftsführung, TMF – Technologie- und Methodenplattform für die vernetzte medizinische Forschung e.V., Berlin, Germany
22 Institut für Sozialmedizin und Gesundheitssystemforschung, Otto-von-Guericke-Universitat Magdeburg Medizinische Fakultat, Magdeburg, Germany
23 Verband Medizinischer Fakultäten, Medizinischer Fakultätentag, Berlin, Germany
24 Stabsstelle Externe Vernetzung und Strategische Kooperation, Charité Universitätsmedizin Berlin, Berlin, Germany

Keywords

healthcare-related data, data integration, Network University Medicine, (NUM), Data Intake and Processing Unit for External Data, (DAAeD), record linkage, health services research

Schlüsselwörter

versorgungsnahe Daten, Datenintegration, Netzwerk Universitätsmedizin, (NUM), Datenannahme- und Aufbereitungsstelle für externe Daten, (DAAeD), Record Linkage, Versorgungsforschung

received 11.11.2025
revised 14.11.2025
accepted 18.11.2025
published online 2026

Bibliography
Gesundheitswesen
DOI 10.1055/a-2780-8325
ISSN 0941-3790

Georg Thieme Verlag KG, Oswald-Hesse-Straße 50, 70469 Stuttgart, Germany

Correspondence
Prof. Wolfgang Hoffmann
Universitätsmedizin Greifswald, Institut für Community Medicine
Ellernholzstr. 1-2
17487 Greifswald
Germany
wolfgang.hoffmann@uni-greifswald.de

ABSTRACT

Linking project data and data from routine clinical practice with healthcare-related data is essential for German healthcare research in order to answer complex questions validly and efficiently. Currently, fragmented data sources, heterogeneous legal requirements, and a lack of infrastructure prevent the optimal use and linking of these data. The Network University Medicine (NUM) is therefore developing a dedicated data infrastructure to link primary study data and routine clinical data with external healthcare-related data (e. g., data from statutory and private health insurance companies, data from cancer registries of the federal states, and data from registration offices). The position paper describes the various data worlds, including examples from epidemiological and clinical research that illustrate the added value and challenges of data linkage. In particular, it presents the new structures of the data acceptance and processing center (DAAeD) in the NUM, which is intended to enable quality-assured, data protection-compliant receipt and linkage of data. Standardized application and approval procedures as well as innovative privacy-preserving record linkage procedures are central to this. International experience, including from Scandinavia and the UK, demonstrates the benefits of such infrastructures for research and healthcare. In conclusion, we advocate rapid political and institutional implementation of the recommendations described in order to make health research in Germany competitive in international comparison and to ensure sustainable, patient-centered health care.

ZUSAMMENFASSUNG

Die Verknüpfung von Projektdaten und Daten aus der klinischen Routine mit versorgungsnahen Daten ist für die deutsche Versorgungsforschung essentiell, um komplexe Fragestellungen valide und effizient zu beantworten. Aktuell stehen fragmentierte Datenquellen, heterogene rechtliche Vorgaben und fehlende Infrastrukturen einer optimalen Nutzung und Verknüpfung dieser Daten entgegen. Das Netzwerk Universitätsmedizin (NUM) entwickelt daher eine darauf ausgerichtete Dateninfrastruktur zur Verknüpfung primärer Studiendaten und klinischer Routinedaten mit externen versorgungsnahen Daten (z. B. Daten der gesetzlichen und privaten Krankenversicherungen, Daten der Landeskrebsregister und Daten der Melderegister). Das Positionspapier beschreibt die verschiedenen Datenwelten einschließlich Beispiele aus epidemiologischer und klinischer Forschung, die die Mehrwerte und Herausforderungen der Datenverknüpfung abbilden. Insbesondere werden die neuen Strukturen der Datenannahme- und -aufbereitungsstelle (DAAeD) im NUM dargestellt, die eine qualitätsgesicherte, datenschutzkonforme Annahme und Verknüpfung von Daten ermöglichen soll. Zentral sind vereinheitlichte Antrags- und Genehmigungsverfahren sowie innovative Privacy-Preserving-Record-Linkage-Verfahren. Internationale Erfahrungen, u. a. aus Skandinavien und UK, belegen den Nutzen solcher Infrastrukturen für Forschung und Versorgung. Es wird für eine rasche politische und institutionelle Umsetzung der beschriebenen Empfehlungen plädiert, um die Gesundheitsforschung in Deutschland im internationalen Vergleich wettbewerbsfähig zu machen und eine nachhaltige, patientenzentrierte Gesundheitsversorgung zu sichern.

Policy Brief

Use and integration of external healthcare-related data in the University Medicine Network (NUM) for German health research

The consistent linking of healthcare-related data from external sources (e. g., data from statutory and private health insurance companies, data from cancer registries of the federal states) with primary study data and data from routine clinical practice is essential in order to answer complex research questions in health services research in a valid and efficient manner. Currently, fragmented data landscapes, heterogeneous legal frameworks, and a lack of infrastructure in Germany hinder the optimal use and linking of these data sources. The University Medicine Network (NUM) is creating a data infrastructure for comprehensive data linking from a variety of sources and their sustainable use for research.

Background

Modern health research requires comprehensive and linkable data sources to evaluate healthcare practices, the effectiveness of new interventions, health economics, and patient safety. This means that the need for data far exceeds the possibilities offered by the Health Data Use Act (GDNG) and the recently launched Health Research Data Center (FDZ).

The German Council of Health and Care Experts (Sachverständigenrat Gesundheit und Pflege, SVR) has repeatedly called for an integrative research data infrastructure with enabling data protection as a prerequisite for the further development of the healthcare system. Other European countries are decades ahead of Germany: Scandinavian countries, the Netherlands, and the United Kingdom are demonstrating how registry data, electronic patient records, and external healthcare-related data can be merged in compliance with data protection regulations and used for population-wide research.

Challenges

Legal requirements and complex application procedures significantly delay the use of linked study, healthcare, and administrative data. Different regulations for the release of data, for example by statutory health insurance companies, as well as heterogeneous conditions at cancer registries of the federal states, make data linkage difficult. The currently inadequate integration of external data sources leads to limited significance and representativeness of study results.

Recommendations for action

- Rapid establishment of a transparent, data protection-compliant infrastructure for data integration. The development and implementation of a standardized data intake and processing unit (DAAeD) in the University Medicine Network (NUM) must be accelerated and secured in the long term.
- Harmonization of the legal framework and approval processes. Uniform application and approval procedures as well as model contracts with data holders should significantly facilitate the use of external healthcare-related data.
- Promotion of technical and organizational solutions for privacy-preserving record linkage. The use of innovative methods for linking data sources in compliance with data protection regulations must be supported and secured by legal regulations.
- Ensuring European integration and connectivity. The future German research data infrastructure in NUM should be developed in line with the best practices of European countries in order to promote international cooperation and research quality.

Political relevance

The targeted promotion of data integration and linkage in the NUM is a decisive lever for improving the quality, results, and efficiency of medical care as well as the innovative capacity of Germany as a research location. A sustainable policy for the use of healthcare-related data ensures patient-centeredness and cost-effectiveness in the healthcare system.

The rapid implementation of the recommendations requires the joint efforts of all relevant stakeholders—politicians, researchers, data holders, and patient representatives—under the guiding principle of a responsible data culture and transparent data management. We can no longer afford not to use the available data.

The NUM University Medicine Network’s goal is making health research in Germany internationally competitive again.

Introduction

In view of the major structural, financial, and social challenges of the German healthcare system, modern health research is faced with the task of answering increasingly complex questions quickly, validly, and reliably based on the best available data [1].

This requires the establishment of a powerful nationwide infrastructure for the effective and efficient conduct of large multicenter studies. At the beginning of the SARS-CoV-2 pandemic, the University Medicine Network (NUM) was founded in 2020 to bundle the research activities of German university medicine and thus enable a faster and more coordinated response to health crises. Within the NUM, all university medical centers in Germany and other cooperating institutions are working to establish a joint research infrastructure for conducting multicenter, prospective, and retrospective clinical and clinical-epidemiological studies and cohort studies.

This research infrastructure offers significant added value when, in addition to the "traditional" primary data collected specifically for a specific study, healthcare-related data that is already routinely generated in the healthcare system as part of treatment or administration is consistently accessed, used, and linked to specific projects.

Until a few years ago, there was no direct research access to treatment data from medical care, in particular to clinical data from university locations. In 2016, the Medical Informatics Initiative (MII) was launched, funded by the then Federal Ministry of Education and Research (BMBF). Since its inception, many researchers and stakeholders have been working to open up access to routine clinical care data for research purposes. The data infrastructure established in the MII currently enables access to routine clinical care data in all German university hospitals and a growing number of cooperating institutions. This data is extracted from the respective hospital documentation by a data integration center (DIZ) at each location, processed, integrated, and made available in a standardized format as a modular core data set. This comprises basic and extension modules that are continuously supplemented [2]. At the beginning of 2023, the DIZs were transferred from MII funding to the NUM. As a result, the routine clinical data from the NUM institutions contained in the core data set are available for research projects on the basis of broad consent (NUM-internal healthcare-related data).

For complex questions regarding the effectiveness, quality, safety, acceptance, and regional differences in diagnostics, treatment, infrastructural, or organizational interventions in the healthcare system, it must be possible to link primary study data and NUM-internal healthcare-related data with additional complementary healthcare-related data (NUM-external healthcare-related data) at the individual level [3–5]. This NUM-external healthcare-related data includes all data relating to healthcare or patients from data repositories that exist outside the NUM. This includes, in particular, billing data from statutory and private health insurance companies, data from disease-related registers, e. g., data from cancer registers of the federal states, and data from residents’ registration offices, in particular vital status information.

Only the linkage of different data sources at the individual level can provide a comprehensive picture of healthcare. The German Council of Health and Care Experts has repeatedly pointed out, most recently in its 2025 annual report, that enabling data protection and an integrative research data infrastructure are necessary prerequisites for the further development of healthcare in Germany, for Germany as a research location, and also for Germany's attractiveness to the research-based pharmaceutical industry [6].

The NUM aims to build and validate this database of primary data, NUM-internal and NUM-external healthcare-related data, not only in isolation for a specific issue, but also to make it permanently accessible for multiple and subsequent use in a large number of research projects with high potential for improving the quality and cost-effectiveness of healthcare in Germany. The needs that NUM covers thus go far beyond the capabilities of the recently launched Research Data Center (FDZ).

The infrastructure of the NUM is geared towards this challenge. Since July 2025 (start of the NUM 3.0 funding phase), the newly designed data intake and processing unit for NUM-external healthcare-related data (DAAeD) is a subproject of the NUM Clinical Epidemiology and Study Platform (NUKLEUS). In the future, the DAAeD will take on the role of accessing, processing, and linking healthcare-related data sources external to NUM—in particular data from health insurance companies, cancer registries and residents' registration offices – to make them usable.

This paper describes the potential of linking healthcare-related data from internal and external NUM sources with primary data from scientific studies based on pressing research questions. First, the various data worlds are described. Examples of studies and questions that can be answered with data from only one of the data worlds are given. The examples are used to explain the strengths and limitations of the individual data worlds. This is followed by an overview of conditions for linking of data from different data worlds. Finally, the NUM-DAAeD structure for providing external healthcare-related data and its linkage is described.

Data worlds in NUM

Primary study data

Primary study data are personal data, including clinical measurements, imaging, and laboratory results from the analysis of biosamples, which are collected in scientific studies (e. g., clinical studies, health services research studies, epidemiological studies such as cohort studies or observational studies) explicitly for scientific use to answer one or more predefined specific research questions. Examples include:

- Randomized clinical trials of new drugs or therapies
- Health services research studies to evaluate and implement a new healthcare model
- Long-term epidemiological studies to investigate disease risks and progression (e. g., disease-related registries or population-based cohorts such as the NAKO Health Study [https://nako.de/])

The strengths of primary study data are high data quality, flexibility in collection, and, in most cases, great data depth determined by the study design, standardized collection, and quality assurance during data collection. In most cases, longitudinal follow-up is possible. Limitations include the often limited scope of the data, the high cost and effort involved in collection, and possible limitations in the representativeness of the database and thus also of the study results.

These factors can limit the possibilities for secondary use of the primary study data. In addition, it is a prerequisite that the informed consent allows further use of the data.

The NUM studies to date have collected particularly extensive data on the quality of biological samples and the resulting molecular data (including -omics). Molecular data generally provide important insights into the development, progression, treatment, and diagnosis of diseases and are of particular importance in medical research. For example, the NUM NAPKON (National Pandemic Cohort Network) cohorts have been extensively characterized at the molecular level. This "biospecimen data" is stored in the NUM's laboratory information management system (LIMS), and participant consent allows it to be linked at any time with other data available in the NUM.

Routine data from clinical care (NUM-internal healthcare-related data)

Routine data from clinical care is data collected in the course of diagnosing and treating patients at the university hospitals and cooperating hospitals participating in NUM. Examples include data from clinical and social histories, diagnoses including severity levels, therapies, laboratory values, imaging findings, and medication data. Since this data reflects actual medical care, it is suitable for answering research questions about care practices in relation to clinical symptoms and findings. Examples of studies using routine clinical data include:

- Studies on processes and patient pathways in hospitals,
- Studies on medication use in hospitals,
- Studies on the care of geriatric patients at the transition between hospital and primary care
- Studies on patient safety, e. g., on drug-related problems and complications

The most important strength of this data is that it reflects real-world medical care. One of the limitations of routine clinical data is that the data can have heterogeneous meanings due to often non-standardized collection and classification. As a result, the data is often only comparable to a limited extent between different institutions. Data quality is often limited by inconsistent documentation, missing values, and systematic biases.

The NUM provides data from routine clinical care at university hospitals and their partner institutions in a standardized output format as a modular core data set. University hospitals account for 10 % of hospital stays in Germany [7]. In the future, data from other hospitals and research practice networks will also be included. It would also be desirable to use data from electronic patient records (ePA) as well as data collected by patients themselves or contributed via health apps or wearables, such as data on quality of life, symptoms, and function, as well as their treatment experiences.

External care-related data (NUM-external healthcare-related data)

External healthcare-related data is crucial for use in the NUM. It is collected outside the NUM as part of administrative processes and in registers. Examples include:

- Billing data from statutory and private health insurance companies

- Data from legally based disease-related registers, in particular cancer registers of the federal states pursuant to Paragraph 65c of the German Social Code, Book V
- Data from voluntary medical registers
- Data from statutory quality assurance, in particular cross-institutional, data-supported quality assurance
- Vital status and address data (residential history) from the residents' registration offices
- Routine monitoring and surveys of the environment, e. g., air quality, noise, climate data

External healthcare-related data can be used to answer questions about healthcare practices, prevalence, distribution, and temporal trends of risk factors and diseases, often without direct reference to individuals. Examples of studies using healthcare-related data include:

- Health economic analyses, cost-effectiveness, appropriateness, and adequacy of diagnostic, therapeutic, preventive, or rehabilitative measures within the framework of standard care
- Investigation of cross-sector care pathways
- Differentiated analyses of the use of medical services
- Quality and safety of medical care (e. g., adherence to guidelines)
- Effectiveness and safety of drugs and medical devices in the context of standard care

The strengths of external healthcare-related data are the often large scope of different variables from various sectors of the healthcare system, patient-related longitudinal continuity, timeliness, completeness, comparatively high standardization, and validity. Many of these sources are representative with respect to the population (e. g., cancer registries, registration offices, official statistics) or cover a large proportion of the total population (e. g., statutory health insurance billing data). One limitation is the potentially restricted usability in relation to research questions: this data was collected for a specific, mostly administrative purpose. This must be taken into account in analyses [8]. Strict legal requirements for the scientific use of this data usually require extensive application, review, and approval procedures. Data transfer often takes place with considerable delay and to a limited extent.

Current conditions for the use of NUM-external healthcare-related data in the NUM

GKV data

The basic prerequisite for the release and use of (social) data from statutory health insurance funds in accordance with Section 75 of SGB X for the scientific use of social data is the voluntary cooperation agreement of the data holders. Data transfer and use then also requires approval from the supervisory authorities of the respective social insurance carrier in accordance with Section 75 of SGB X.

Individual health insurance companies also offer the evaluation of anonymized data either in their own data warehouse or as a data export without a §75 application; in both cases, linking to other data sources is not possible.

Since October 2025, the Health Research Data Center (FDZ), located at the Federal Institute for Drugs and Medical Devices (BfArM), has been offering data from all statutory health insurance companies for analysis on the basis of SGB V §303a-f (https://www.forschungsdatenzentrum-gesundheit.de/aktuelles). In the future, there are plans to link this data with data from the ePA and cancer registries. Use of this data is currently planned within a so-called Trusted Research Environment within the FDZ. There are currently no plans to pass this data on to researchers, for example for linking with study data or NUM-internal care-related data, as envisaged by the NUM [9].

Cancer registry data

Nationwide consolidated cancer registry data can be requested from the Center for Cancer Registry Data at the Robert-Koch-Institute (RKI). However, this data cannot be linked to other data sources, which is why this case will not be considered further. The data from the cancer registries in accordance with §65c SGB V can also be released by the individual cancer registries of the federal states upon request; each cancer registry has its own use and access procedure for this purpose. Since 2025, a standardized nationwide application form can be used to submit uniform requests for data release to the cancer registries of several federal states. However, the data usage agreements and the conditions for data usage specified therein continue to be regulated and organized differently between the federal states.

Connecting the data worlds

Answering complex research questions usually requires information from multiple data sources [10]. This is also a necessary prerequisite for meaningful analyses of correlations between certain individual characteristics, risk factors, exposures, and specific health-related endpoints and costs. Plausibility checks and mutual compensation for missing variables in the individual data types also often require personal data to be merged. The resulting integrated data sets can be used to specify target groups, relevant influencing factors, and potential confounders, as well as to estimate the number of cases required and provide further information for planning interventional studies. Last but not least, well-designed analyses based on person-related linked data sets can provide clues to successful recruitment strategies. Specific examples of research topics that can be addressed by linking external healthcare-related data with primary or routine clinical data, including imaging and laboratory diagnostics, are:

- Clinical research: extension of follow-up periods, mapping of long-term outcomes and rare side effects and complications
- Personalized medicine: Identification of risk groups, evaluation of the effect of tailored therapies, also in comparison to conventional therapies, prediction of therapy response
- Population health: Effects of environmental factors on health, identification and characterization of health inequalities, differentiated analysis of healthcare system utilization and resulting costs
- Development of preventive measures, review of the effectiveness and cost-efficiency of implemented preventive measures at the population level or in defined risk groups
- Epidemiological research: Early detection of disease outbreaks, analysis of risk factors, assessment of the effectiveness of vaccination programs, evaluation of screening measures

- Health services research: Effectiveness and health economic efficiency of complex healthcare interventions
- Patient safety: Identification of side effects and complications of therapies (e. g., drugs, surgical procedures, radiation therapy, or combinations thereof), including nosocomial infections and multidrug-resistant pathogens
- Rare diseases: Identification of patients, development of therapies, improvement of quality of life
- Resilience of the healthcare system: Analysis of determinants and requirements for a resilient healthcare system; infection epidemiology; monitoring of quality and safety, effectiveness, and access to measures in connection with indicators for infection prevention and control.

Linking NUM-external care-related data with primary data and NUM data from routine clinical care

Linking NUM-external healthcare-related data with primary data or data from routine clinical care (record linkage) is still a challenge in many cases today and often not possible [9]. As a rule, the consent of the study participants is required as a legal basis, as well as the approval of the responsible ethics committee. However, both are often not sufficient at present. In the case of statutory health insurance funds, for example, the approval of the respective supervisory authorities is also required for the scientific use of social data. And even if the study participants have given their informed consent, the statutory health insurance funds are not yet obliged to participate in the procedure. This can be interpreted as a barrier to the informational self-determination which is guaranteed by the German Constitution.

If data from cancer registries is required, approval from the relevant cancer registry committees is necessary. For individual linkage of cancer registry data with other data sources, the consent of the test subjects is usually required. In most cases, the linkage is carried out with the involvement of a Trusted Third Party on the basis of identifying data that is available in all data sets to be linked. Examples of this are the insured person's ID or health insurance number, or even name, place of residence, and date of birth. The data providers only ever see the identifying data of their own insured persons or patients (privacy-preserving procedure). The researchers ultimately receive pseudonymized data sets for analysis. Where it is possible to link data sets without the consent of the test subjects/patients on the basis of a specific legal basis, special privacy-preserving procedures are used to prevent the unauthorized identification of individual test subjects/patients.

The current draft bill for a Medical Register Act [11] takes an important step by creating a legal basis for storing the unchangeable part of the health insurance number (KVNR) for qualified medical registers.

The Data Intake and Processing Unit for External Data (DAAeD) in the University Medicine Network (NUM)

The NUM aims to significantly improve the currently completely unsatisfactory situation with regard to data linkage. To this end, a Data Intake and Processing Unit for external healthcare-related data (NUM DAAeD) is being developed and implemented. The DAAeD is intended to enable the large-scale use of high-quality external healthcare-related data on the basis of informed consent from patients [10].

The specific objectives of the DAAeD are:

- Secure and standardized intake of healthcare-related data from external data sources
- Plausibility checking, harmonization, and preparation of data with a focus on data quality and integrity
- Linking of primary project data and NUM-internal healthcare-related data with NUM-external healthcare-related data
- Creation of indicators for research purposes
- Provision of data in a format suitable for research
- Compliance with data protection regulations and ethical standards

A number of functions and services are required to achieve these objectives:

- Receipt and cross-checking of informed consents from study centers and DIZs
- Informing data holders about the availability of informed consent forms
- Data intake and validation
- Data cleansing and transformation
- Data harmonization and standardization
- Linking of data by a Trusted Third Party
- Data storage and provision of linked data
- Data anonymization or pseudonymization
- Transfer of data to users via the NUM transfer point
- Extraction and cataloging of metadata

Close cooperation within the framework of the PLATO 2 project is being sought for the use of data from cancer registries. PLATO 2 is a concept for creating a platform that enables nationwide, event-driven data consolidation and analysis of cancer registry data from the federal states, as well as linking cancer registry data with other data. This is intended to implement the so-called second stage of cancer registry data consolidation regulated in the Act on the Consolidation of Cancer Registry Data (Paragraph 10, sentence 2 of the Federal Cancer Registry Data Act (BKRG)) [12].

Researchers receive advice and support from the NUM Methods and Biospecimen Hub (NUM-MB) for all steps in the design, planning, and preparation of a research project based on data linkage.

Examples of studies and research questions

The following section presents some examples of relevant research questions, highlights key challenges, and explains possible solutions involving NUM-DAAeD and, in the case of oncological questions, PLATO 2.

Example 1: Influence of immune checkpoint therapy in malignant melanoma – study concept for real-world care

The treatment of metastatic malignant melanoma has improved significantly thanks to immune checkpoint inhibitors (ICIs), but this is accompanied by increased immune-related adverse events (irAEs). This study examines the effects of ICI therapy on complete remission, survival, and quality of life of patients at the population level. The focus is on the duration of therapy, the sectoral allocation of treatment (who, what, how prescribed), and the occurrence and management of side effects in order to evaluate drug safety (pharmacovigilance).

Key research questions:

- Do specialized centers improve clinical outcomes? How great is the patient benefit of treatment in specialized centers compared to the current standard?
- How are irAEs currently treated and which strategies are promising?
- How common are severe side effects, and how often do they lead to hospitalizations?
- What are the predictors of severe side effects? What strategies prevent the occurrence of severe side effects?
- Are there correlations between treatment success and side effect management?
- Are there regional differences in therapy implementation?

Linking data sources:

This example combines healthcare data from statutory health insurances, data from cancer registries in all federal states, patient surveys on symptom burden and quality of life, and other data sources (structured quality reports from hospitals, certification data from the German Cancer Society (DKG), indicators and maps for spatial and urban development (INKAR), etc.). Linking these data sets could enable a comprehensive view of the effects of therapy, as well as the risks and side effects and their management in healthcare practice. This approach allows for a comprehensive analysis of influencing factors and therapy outcomes and the identification of regional differences.

Special feature:

This example highlights the need for interdisciplinary collaboration between methodological experts and oncologists in order to correctly interpret clinical correlations and achieve evidence-based improvements in care. The use of NUM and PLATO 2 structures is intended to facilitate data integration and analysis [12].

Example 2: DigiNet: A prospective comparative cohort study to optimize and evaluate digitally networked and personalized care for patients with advanced non-small cell lung cancer (NSCLC)

DigiNet aims to optimize personalized care for patients with stage IV non-small cell lung cancer (NSCLC). To this end, healthcare providers across the spectrum of care are digitally networked with specialized centers. The project builds on the structures of the National Network for Genomic Medicine in Lung Cancer (nNGM). The DigiNet intervention cohort is actively recruited in participating clinics and specialist practices, while the population-based comparison cohort is formed by matching cohorts within the cancer registries of the study regions of North Rhine-Westphalia, Saxony, and Berlin. The control group comprises stage IV NSCLC patients who are not participating in DigiNet and/or the nNGM.

Key areas of evaluation:

1. Healthcare evaluation at the patient level: survival, use of therapy, patient-reported endpoints
2. Health economic evaluation
3. Formative evaluation (qualitative process analysis)

Linking of data sources:

For the DigiNet intervention cohort, primary data is linked to health insurance billing data on an individual basis. For the population-based comparison cohort, cancer registry data is also linked to health insurance billing data on an individual basis. On this basis, the healthcare and health economic evaluations could be carried out. Since no individual informed consent was available for the population-based comparison cohort, a comprehensive concept for privacy-preserving record linkage (PPRL) of cancer registry and billing data based on health insurance numbers was developed and implemented [13].

Special feature:

In addition to the primary study data, cancer registry data and health insurance billing data are integrated into the project. DigiNet thus goes beyond classic clinical studies and pursues a multi-source data integration approach that combines the strengths of different data sources. Only by including healthcare-related data, it is possible to make a valid and robust assessment of the new form of healthcare in terms of overall survival and cost-effectiveness at the population level in the study area. At the same time, DigiNet highlights the potential of cancer registries for generating population-based comparative cohorts and their central importance for the evidence-based evaluation of the effectiveness and benefits of new models of care [14].

Example 3: RubiN: A prospective, comparative cohort study is investigating whether the implementation of cross-sector, multi-professional care and case management (CCM) for geriatric patients can close gaps in care, improve patients' autonomy in their home environment, and enable them to live independently in their own homes.

Not all older people are geriatric patients. Geriatric patients often suffer from acute and chronic illnesses, sometimes simultaneously, and these are often difficult to distinguish from the physiological processes of aging. Functions such as walking, seeing, hearing, and continence are often impaired, which frequently prevents patients from remaining in their own homes. In the project, care and case managers used assessments to record the care needs of geriatric patients in detail. Based on this, individual treatment and care pathways were developed and implemented for the patients in the intervention group, while the control group received "care as usual."

Key areas of evaluation:

1. Healthcare evaluation: Ability to cope with everyday activities based on primary data collected at various measurement points
2. Health economic evaluation: Cost-effectiveness analysis from the perspective of the statutory health insurance fund (cost-utility analysis)
3. Qualitative and quantitative process evaluation: Feasibility, acceptance, satisfaction

Linking of data sources:

In addition to the primary study data, health insurance billing data was used to answer the research questions in the RubiN project. The billing data enables an analysis of the use of health services and, based on this, a health economic evaluation of patient care. Also based on healthcare-related data, the impact of the RubiN intervention on the quality of outpatient care could be assessed using outpatient-sensitive hospital admissions.

Special feature:

The personal linking of primary study data with health insurance billing data enables a well-founded and comprehensive evaluation of the CCM care model, taking into account both qualitative and quantitative aspects [15].

International perspective

The successful use of record linkage for research has been established in several European countries for many years. The Scandinavian countries and the UK in particular show that medical registries, electronic health records, and care-related data can be efficiently merged while adequately complying with data protection standards. Under the framework conditions of the General Data Protection Regulation (GDPR), sustainable structures have been created in these countries that enable research and care data to be used for the evaluation of clinically relevant endpoints such as mortality, quality of life, side effects, or therapy efficiency.

A key success factor is the establishment of technical and organizational infrastructures that ensure secure, standardized, and traceable data linking. While Sweden and Finland use national identification numbers, the United Kingdom, among others, uses Trusted Research Environments (TRE) as secure analysis platforms. In Denmark, the national registry system allows the prospective use of linked health data for population-wide research questions. These models show that data protection, research, and patient interests are compatible when governance structures, ethical oversight, and transparency are considered together. Denmark and Finland demonstrate that a legally anchored, centrally coordinated data infrastructure benefits researchers and policymakers alike – because it produces robust, reproducible evidence with immediate benefits for patients. For Germany, this presents an opportunity to strategically develop existing approaches (DAAeD, NUKLEUS) within the framework of the NUM and embed them in a structure that is compatible with Europe, as required by the Common European Data Spaces.

Discussion and outlook

The use and linking of data from different sources, especially healthcare-related data, can significantly improve the informative value of research projects or, in many cases, enable research questions to be answered in the first place. Important complex research questions regarding the effectiveness of an existing or new intervention and the associated costs and risks can only be answered on the basis of the personal linking of several healthcare-related data sources. This requires structures that make this data available in high quality and advise and support researchers in its use.

The NUM DAAeD is a necessary prerequisite for the sustainable development of this research. This requires robust and transparent governance structures, processes, and contracts with the data holders. There are still considerable challenges to be overcome, e. g., the procedures must be implemented in such a way that all ethical standards are complied with and data protection and regulatory issues need to be further clarified. A responsible data culture and transparent data management at every stage of the study are required. By linking primary research data with NUM-internal and NUM-external healthcare-related data on the basis of informed consent from study participants, the quality, results, and efficiency of medical care can be considered together. This makes it possible to identify where care provides the greatest benefit and creates a basis for value-oriented, patient-centered, and at the same time economically sustainable healthcare.

We can no longer afford not to use the available data.

Conflict of Interest

The authors declare that they have no conflict of interest.

References

[1] Hoffmann W, Härter M, Bierbaum T et al. Position paper of the German Network for Health Care Research: What is health care research? Gesundheitswesen 2025 English, German. DOI: 10.1055/a-2592-5998. Epub ahead of print PMID: 40876827

[2] Ammon D, Kurscheidt M, Buckow K et al. Arbeitsgruppe Interoperabilität: Kerndatensatz und Informationssysteme für Integration und Austausch von Daten in der Medizininformatik-Initiative [Interoperability Working Group: core dataset and information systems for data integration and data exchange in the Medical Informatics Initiative]. Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz 2024; 67: 656–667. German. DOI: 10.1007/s00103-024-03888-4. Epub 2024 May 16 PMID: 38753022 PMID: PMC11166738

[3] Klinkhammer-Schalke M, Kaiser T, Apfelbacher C et al. Manual für Methoden und Nutzung versorgungsnaher Daten zur Wissensgenerierung [Manual for Methods and Use of Routine Practice Data for Knowledge Generation]. Gesundheitswesen 2020; 82: 716–722. German. DOI: 10.1055/a-1237-4011. Epub 2020 Sep 22 PMID: 32961567 PMID: PMC7661315

[4] Hoffmann F, Kaiser T, Apfelbacher C et al. Versorgungsnahe Daten zur Evaluation von Interventionseffekten: Teil 2 des Manuals [Routine Practice Data for Evaluating Intervention Effects: Part 2 of the Manual]. Gesundheitswesen 2021; 83: e40. German. DOI: 10.1055/a-1516-8823. Epub 2021 Jun 2. Erratum for: Gesundheitswesen. 2021 Jun;83(6):470-480. doi: 10.1055/a-1484-7235 PMID: 34077984

[5] Veit C, Bierbaum T, Wesselmann S et al. Versorgungsnahe Daten für Versorgungsanalysen – Teil 3 des Manuals [Routine Practice Data for Health Care Analyses: Part 3 of the Manual]. Gesundheitswesen 2023; 85: 718–724. German. DOI: 10.1055/a-1966-0104. Epub 2022 Dec 19 PMID: 36535653 PMID: PMC10444518

[6] Sachverständigenrat Gesundheit und Pflege. Gutachten 2025 "Preise innovativer Arzneimittel in einem lernenden Gesundheitssystem". Zugriff 02.11.2025 https://www.svr-gesundheit.de/publikationen/gutachten-2025/

[7] Verband der Universitätsklinika Deutschlands. Zugriff 04.11.2025 https://www.uniklinika.de/die-deutschenuniversitaetsklinika/zahlen-und-fakten/

[8] Swart E, Gothe H, Geyer S et al.German Society for Social Medicine and Prevention; German Society for Epidemiology Gute Praxis Sekundärdatenanalyse (GPS): Leitlinien und Empfehlungen. Gesundheitswesen 2015; 77: 120–126

[9] Schmitt J, Ihle P, Hoffmann W et al. 2023; Gutachten im Auftrag des Bundesministeriums für Gesundheit (BMG): Record Linkage der Daten des FDZ Gesundheit mit extern gehaltenen Forschungsdaten anhand des Beispiels einer Verknüpfung mit Daten einer Kohorte des NAPKON. https://www.uniklinikum-dresden.de/de/das-klinikum/universitaetscentren/zegv/news/GA_NAPKON_FDZ_V2_1_Zusammenfassung.pdf

[10] Schmitt J, Ihle P, Schoffer O et al. und die gemeinsame Arbeitsgruppe „Externe Daten" des Netzwerks Universitätsmedizin (NUM) und der Medizininformatik-Initiative*Datennutzung für eine bessere Gesundheitsversorgung–Plädoyer für eine kooperative Forschungsdatenplattform der gesetzlichen und privaten Krankenversicherung und dem Netzwerk Universitätsmedizin (NUM) [Access to and use of Data for better Healthcare: A Plea for a cooperative data and Research Infrastructure of Statutory and Private Health Insurers and the Network University Medicine (NUM)]. Gesundheitswesen. 2024 German. DOI: 10.1055/a-2438-0670 Epub ahead of print. Erratum in: Gesundheitswesen. 2025 Jan 31. doi: 10.1055/a-2501-9737 PMID: 39389575

[11] Aktueller Referentenentwurf für ein Medizinregistergesetz https://www.bundesgesundheitsministerium.de/fileadmin/Dateien/3_Downloads/Gesetze_und_Verordnungen/GuV/M/Medizinregistergesetz_RefE.pdf Zugriff 04.11.2025

[12] Konzept zur Schaffung einer Plattform für eine bundesweite anlassbezogene Datenzusammenführung und -analyse in der Onkologie https://www.bundesgesundheitsministerium.de/service/publikationen/details/plato-2.html Zugegriffen: 21.10.2025

[13] Kästner A, van den Berg N, Kron A et al. Nutzung und Verknüpfung versorgungsnaher Daten zur Evaluation neuer Versorgungsformen am Beispiel des Innovationsfondsprojekts „DigiNet". 23. Deutscher Kongress für Versorgungsforschung (DKVF). Potsdam, 25.-27.09.2024.Düsseldorf: German Medical Science GMS Publishing House; 2024 Doc24dkvf288. DOI: 10.3205/24dkvf288

[14] Kästner A, Kron A, Eilers L et al. DigiNet: Optimizing personalized care for patients with stage IV non-small cell lung cancer (NSCLC) through a digitally connected provider network-analysis plan of a prospective multicenter cohort trial. J Cancer Res Clin Oncol 2025; 151: 244. DOI: 10.1007/s00432-025-06275-x. PMID: 40924208 PMID: PMC12420542

[15] Gloystein S, Thomé F, Goetz K et al. RubiN – continuous care in regional networks: a study protocol for a prospective controlled trial. BMC Geriatr 2021; 21: 183. DOI: 10.1186/s12877-021-02106-z. PMID: 33726695 PMID: PMC7962254

This article is part of the DNVF Special Iissue "Health Care Research and Implementation"