How Researchers Access Clinical Data: Resources, Pathways, and Challenges

Silvia Becker[1,2,*], Lea Sophie Singson[3], Ely Nova Natalia Silaban[3], Ajay Krishna[2], Till Keller[2,4], Dirk Westermann[2], Thomas Arentz[2], Franziska Boehm[3], Martin Eichenlaub[2], and Axel Loewe[1]

[1]Karlsruhe Institute of Technology, Institute of Biomedical Engineering, Karlsruhe, Germany
[2]Department of Cardiology and Angiology, University of Freiburg, Bad Krozingen, Germany
[3]FIZ Karlsruhe - Leibniz Institute for Information Infrastructure, Eggenstein-Leopoldshafen, Germany
[4]Department of Cardiology, Justus Liebig University Giessen, Giessen, Germany
[*]Corresponding author: publications@ibt.kit.edu.

Introduction

Digitalisation increasingly influences all areas of research, particularly data management processes such as data generation, accessibility, and analysis. In the health sector, digitalisation has led to the rapid growth of digital health data. For instance, the digital acquisition of electrocardiograms (ECGs) resulted in a vast amount of data, exploitable for biomedical research.

Objective

This work aims to provide an overview of the various sources of clinical data used in research, their access pathways and bottlenecks encountered throughout the data acquisition and utilisation process by researchers and their institutions. While the analysis focuses on the German context, many of the structural challenges, workflows, and constraints described here are shared across healthcare systems, especially in the European Union, due to data protection rules.

Methods

Figure 1 provides an overview of three different pathways for accessing clinical data. These pathways are described in detail below.

Case 1: Data access through employment

In this scenario, researchers access clinical data through the hospital network, as employees are typically granted access to relevant internal data systems.

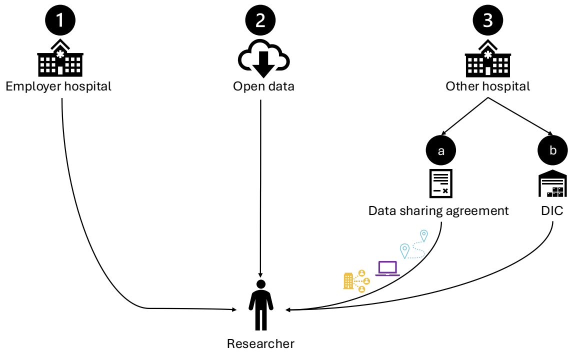

Figure 1: Schematic overview of clinical data sources and access pathways for researchers. Case 1 illustrates data access as a hospital employee, Case 2 through open data, and Case 3 via another hospital. In Case 3, two access pathways are shown: through a data sharing agreement or a Data Integration Centre (DIC). Coloured icons indicate different access modalities, such as virtual private network (VPN) access, direct data download, on-site data use, or indirect data use within a federated learning setup.

Case 2: Data access via open databases

Open-access platforms such as PhysioNet [1] host anonymised clinical datasets that are freely available to researchers. Some datasets can be accessed without additional requirements, while others may require registration or proof of ethical training.

Case 3: Data access through another hospital

Accessing clinical data from another hospital can occur in two main ways: (a) through a data sharing agreement or (b) via a Data Integration Centre (DIC). In case (a), the mode of data sharing may vary. Researchers might gain direct access via VPN, download anonymised or pseudonymised data, use the data only on-site at the hospital, or access it indirectly through federated learning, where machine learning models are trained collaboratively without transferring the underlying data [2]. In case (b), the DIC acts as a fiduciary and provides researchers with pseudonymised datasets, thereby facilitating access to large and diverse data collections while maintaining compliance with data protection regulations.

Results

In the scope of the DiTraRe (Digital Transformation of Research) project, we investigated three data access pathways.

Case 1: Data access through employment

Researchers employed by a hospital can access clinical data through the internal network. Access is limited to the institution's data and computing resources. Additional or external datasets require other pathways. This approach typically requires moderate time and minimal additional financial resources if staff are already employed.

Case 2: Data access via open databases

Open-access databases such as PTB-XL [3, 4] and MIMIC-IV [5] provide anonymised clinical datasets that can be accessed directly or after registration and training. This pathway is fast and inexpensive, but may not meet highly specific data or metadata needs. Main bottlenecks include dataset availability, quality, and confidentiality, which can prevent some data from being openly published.

Case 3: Data access through another hospital

Access via another hospital can occur through (a) data-sharing agreements or (b) a Data Integration Centre (DIC). We found that data-sharing agreements, which require ethics approval, often do not provide the most convenient or efficient form of data access, and are further complicated by limited interoperability between systems, which hinders the integration and reuse of datasets.

When access is provided via a Data Integration Centre (DIC), researchers first obtain ethics approval (votum), after which the DIC exports the data in an organised and pseudonymised format. While this allows secure access to larger datasets and fosters interoperability, it is resource-intensive: personnel capacity and communication within DICs can slow processes, and some DICs operate with small teams. The primary bottleneck in this case is therefore the communication flow and available personnel capacity within the respective DICs.

Discussion

Each data access pathway offers distinct advantages and limitations. Employment-based access provides comprehensive internal data at moderate cost but is limited in scalability across institutions. Open databases allow rapid, low-cost research while ensuring privacy compliance, but are constrained by scope, specificity, and dataset quality. Access via other hospitals or Data Integration Centres (DICs) enables external validation and access to larger datasets but involves higher time and personnel demands as well as administrative bottlenecks.

DIC-based access holds significant potential for researchers, particularly when no suitable open dataset exists. In Germany, around 40 DICs have been established. However, the primary bottlenecks are communication flow and available personnel capacity, highlighting the importance of sufficient staffing and organisational support.

To advance data-driven biomedical research, future efforts should focus on simplifying and accelerating ethical and legal procedures. Furthermore, DIC capacities should be strengthened through adequate institutional and governmental support. Finally, promoting interoperability via harmonised data formats enables integration across hospitals and datasets, improving data reuse and analysis.

Clinical data access pathways vary in scope, efficiency, and resource requirements, and recognizing these differences is key to planning and conducting data-driven biomedical research effectively.

References

[1] A. L. Goldberger, L. A. Amaral, L. Glass *et al.*, "PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals," *Circulation*, vol. 101, no. 23, pp. e215–e220, 2000.

[2] C. Zhang, Y. Xie, H. Bai, B. Yu, W. Li, and Y. Gao, "A survey on federated learning," *Knowledge-Based Systems*, vol. 216, p. 106775, 2021.

[3] P. Wagner, N. Strodthoff, R.-D. Bousseljot, D. Kreiseler, F. I. Lunze, W. Samek, and T. Schaeffter, "PTB-XL, a large publicly available electrocardiography dataset," *Scientific Data*, vol. 7, no. 1, pp. 1–15, 2020.

[4] N. Strodthoff, T. Mehari, C. Nagel, P. J. Aston, A. Sundar, C. Graff, J. K. Kanters, W. Haverkamp, O. Dössel, A. Loewe *et al.*, "PTB-XL+, a comprehensive electrocardiographic feature dataset," *Scientific Data*, vol. 10, no. 1, p. 279, 2023.

[5] A. E. Johnson, L. Bulgarelli, L. Shen, A. Gayles, A. Shammout, S. Horng, T. J. Pollard, S. Hao, B. Moody, B. Gow *et al.*, "MIMIC-IV, a freely accessible electronic health record dataset," *Scientific Data*, vol. 10, no. 1, p. 1, 2023.