

12-31-2025

## Data Work in Healthcare. Mediating data quality and data governance in a data-intensive world

Bjarki Freyr F. Sveinbjarnarson  
*Reykjavik University, bjarkis@ru.is*

Lisa Schmitz  
*Reykjavik University, lisas@ru.is*

Erna Sif Arnardottir  
*Reykjavik University, ernasifa@ru.is*

Anna Sigridur Islind  
*Reykjavik University, islind@ru.is*

Follow this and additional works at: <https://aisel.aisnet.org/sjis>

---

### Recommended Citation

Sveinbjarnarson, Bjarki Freyr F.; Schmitz, Lisa; Arnardottir, Erna Sif; and Islind, Anna Sigridur (2025) "Data Work in Healthcare. Mediating data quality and data governance in a data-intensive world," *Scandinavian Journal of Information Systems*: Vol. 37: Iss. 2, Article 15.

DOI: 10.17705/3SJIS/037.24

Available at: <https://aisel.aisnet.org/sjis/vol37/iss2/15>

This material is brought to you by the AIS Journals at AIS Electronic Library (AISeL). It has been accepted for inclusion in Scandinavian Journal of Information Systems by an authorized administrator of AIS Electronic Library (AISeL). For more information, please contact [elibrary@aisnet.org](mailto:elibrary@aisnet.org).

---

## **Data Work in Healthcare. Mediating data quality and data governance in a data-intensive world**

### **Cover Page Footnote**

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 965417

# Data Work in Healthcare

## Mediating data quality and data governance in a data-intensive world

Bjarki Freyr Sveinbjarnarson, Lisa Schmitz, Erna Sif Arnardóttir, & Anna Sigridur Islind

Reykjavik University, Department of Computer Science, Iceland.

*[bjarkis@ru.is](mailto:bjarkis@ru.is), [lisas@ru.is](mailto:lisas@ru.is), [ernasifa@ru.is](mailto:ernasifa@ru.is) & [islind@ru.is](mailto:islind@ru.is)*

**Abstract.** Data is a critical resource in healthcare research, yet ensuring high data quality remains a persistent challenge. Errors introduced at the point of data entry can propagate through the data lifecycle, affecting usability, integrity, and research outcomes. In response to this challenge, we designed and developed a digital infrastructure over the course of four years capable of accommodating diverse structured health data while enforcing predefined data standards through a validation system. We use that digital infrastructure to examine the root causes of data quality issues at the input stage by analysing the type of data work needed for meaningful data curation. Through a combination of feasibility evaluation, surveys, and semi-structured interviews, we identified recurring mistakes, user behaviour patterns, and underlying reasons for poor data quality. We contribute six key elements of data work necessary for meaningful decision-making and research purposes, grouped under three categories: (i) the data work needed for reaching intrinsic data quality in data governance, (ii) the data work needed for reaching contextual data quality in data governance, and (iii) the data work needed for reaching representational and accessibility data quality in data governance.

**Key words:** Data Quality, Data Work in Healthcare, Data Governance, Data Quality, Information Systems, Healthcare Data.

## 1 Introduction

Data has become a central resource in the modern world, yet the effort required to make it usable is often overlooked (Abbasi et al., 2016; Bossen et al., 2019). This issue is particularly evident in healthcare, where data-driven decision-making and personal-

Accepting editor: Claus Bossen

ized medicine are increasingly prioritized. As the demand for high-quality, actionable data grows, so does the need to investigate and document the data work involved in achieving these goals (Davenport & Patil, 2012; Parmiggiani et al., 2022). This work, often described as backstage (Parmiggiani et al., 2022) or invisible (Islind et al., 2021), is gaining recognition in the literature as a crucial yet frequently underestimated aspect of data governance. The digitalization of healthcare, supported by digital infrastructures and various healthcare technologies, aims to improve efficiency and quality of care through data. However, data does not simply exist; it must be produced, collected, structured, validated, and interpreted (Cruz, 2023; Gregory et al., 2019; Su et al., 2022). This effort is known as data work (Bossen et al., 2019), encompassing technological, analytical, and contextual tasks required to make data usable within healthcare (Fiske et al., 2019).

Data work in healthcare is inherently complex, requiring not only technical skills but also domain-specific knowledge to maintain data quality and integrity (Jones, 2019). Data work involves continuous human interaction with data curation processes and technologies, making data work prone to errors and inconsistencies. These inaccuracies can significantly impact healthcare decisions, diagnoses, and patient outcomes, alongside research outcomes down the line if faulty data is used for research purposes. Given these potential consequences, examining the intricacies of data work and the factors contributing to errors is crucial. By understanding where and how inaccuracies arise, better informed strategies for improving the data quality of health data can be formed, but for that to happen, data work needs to be put center stage.

Healthcare relies heavily on rigorous research, which in turn depends on high-quality data. As the volume of data flowing through healthcare systems increases, so do the tasks required to manage, process, and ensure its reliability (Janssen et al., 2016). Understanding the data work underpinning healthcare research is becoming increasingly important, particularly as data-driven digital infrastructures expand. While these data-driven digital infrastructures enhance both the quality and quantity of available data, they also generate a substantial amount of behind-the-scenes work. This backroom data work is essential for shaping raw data into a usable form, enabling researchers and practitioners to leverage its full potential. With the rise of big data and the ongoing process of datafication, data-driven digital infrastructures have become a cornerstone of digital transformation strategies in healthcare (Galliers et al., 2017; Parmiggiani et al., 2022). As a result, multidisciplinary teams now collaborate to discover, transform, model, and report vast and complex datasets for research purposes (Davenport & Patil, 2012; Parmiggiani et al., 2022). This transformation involves both visible and invisible data work, spanning data collection, data analysis, and securing data quality.

Despite the increasing recognition of the need for high-quality data, much of the existing literature focuses on the benefits and potential of digital infrastructures—such as efficiency, improved service quality, and enhanced decision-making—while overlooking the extensive data work required to achieve these outcomes (Bertelsen et al., 2024). Researchers play a critical role in this process, yet the effort involved in preparing, curating, and structuring data remains largely behind the scenes, still invisible to a large extent.

As data-driven digital infrastructures become the backbone of healthcare research, understanding how they shape and are shaped by data work is essential. This paper highlights the often-overlooked efforts of researchers in establishing and maintaining the scaffolding for data-driven healthcare, emphasizing the labour-intensive process of transforming raw data into meaningful, high-quality information by examining the data work in detail. Our research focuses on improving data quality in healthcare by identifying the necessary data work linked to meaningful contextual data quality by developing a lean database structure alongside a rigorous validation system embedded within our digital infrastructure. Our empirical context is within sleep research—a context that, despite its profound impact on health, has historically been overlooked as a major pillar of well-being. Sleep influences hormonal regulation, organ function, cognitive and physical performance, and overall health, making it a crucial factor in daily life (Arnardottir et al., 2021; Schmitz et al., 2025).

Sleep disorders, such as insomnia, which affects 10% of adults, and sleep apnea, which impacts nearly 1 billion people worldwide, are prevalent across numerous research fields (Arnardottir et al., 2021; Hodge et al., 2025). This diversity is reflected in the data collected, spanning multiple disciplines, and presents unique challenges in building and governing heterogeneous data. Our digital infrastructure has accumulated data from various sources, leading to significant data governance complexities. The need to merge smaller studies into larger datasets has further amplified data quality issues, particularly as responsibilities are distributed among researchers in various roles and from multiple disciplines; each with different priorities and approaches. As is common in healthcare research, data collection often centers on individual study objectives rather than the collective and cohesive effort required to ensure high-quality data for future decision-making and research. This paper examines these challenges, highlighting the invisible data work necessary to bridge the gap between data acquisition and the long-term usability of thoughtful and reliable healthcare data.

We designed and developed the aforementioned digital infrastructure on the foundations of a lean database structure, capable of accommodating diverse healthcare data without requiring additional tables, thereby minimizing complexity. It currently hous-

es 63 distinct data sources, encompassing approximately 2,632 parameters from over 2,000 participants, all contributing complex sleep-related data. Over the past four years, researchers working with this data have frequently encountered errors that bypassed the validation system, prompting continuous refinements. To mitigate these challenges, we integrated a robust validation system within our digital infrastructure. In this paper, we examine how effectively researchers can independently upload their collected sleep data using our digital infrastructure and examine the data work connected to acquiring and working with robust research data. Our empirical data is based on a mixed-methods study involving thirteen participants, combining quantitative data with qualitative insights from semi-structured interviews. Our paper poses the following research questions: *What kind of data work is involved in curating quality healthcare data?*

Our analysis identifies three key elements of data work necessary for curating healthcare data for meaningful decision-making and research purposes: (i) the data work needed for reaching intrinsic data quality in data governance (ii) the data work needed for reaching contextual data quality in data governance, and (iii) the data work needed for reaching representational and accessibility data quality in data governance. Moreover, the validation system embedded within our data-driven digital infrastructure proved its effectiveness in enforcing standards, as no participant was able to upload all data sources without the system identifying at least one mistake, helping the participants along in their data work. Our findings emphasize the need to make data work more visible and integrate it as a fundamental, widely recognized aspect of scientific inquiry, and the need to have both systemic measures in place, as well as a cohesive human understanding across teams. We also argue that decisions about data quality and data governance must move beyond behind-the-scenes work and become a visible, integral part of research practice (Barley & Bechky, 1994) or data science (Parmiggiani et al., 2022). Instead, these discussions must take centre stage, ensuring that data governance and data quality standards are explicitly acknowledged and actively incorporated into research efforts where healthcare data is collected, managed, and analysed.

## 2 Related work

A growing discourse in the literature emphasizes adapting data governance to specific contexts to manage the increasing complexity and volume of data. Data governance is a broad concept, generally encompassing data definition, implementation, and monitoring (Alhassan et al., 2016). It also involves establishing policies, standards, and guidelines to define when data quality is considered sufficient (Alhassan et al., 2016). A key aspect of data governance is the assignment of roles, tasks, and responsibilities to en-

sure compliance with these standards (Benfeldt et al., 2020). Additionally, continuous monitoring and iterative adjustments to roles, guidelines, and processes are essential for improving data quality over time (Petzold et al., 2020).

However, implementing data governance models presents challenges such as defining responsibilities, balancing strictness (Janssen et al., 2020; Petzold et al., 2020), and overcoming resistance to change (Ladley, 2019). Furthermore, errors are common during the adoption of data governance, which can have unintended ripple effects, such as increased data work. Effective error prevention is, therefore, crucial in establishing sustainable and efficient data governance frameworks. For any data governance model to be effective, well-defined data processes are essential. The literature on data governance highlights that a lack of shared understanding of these processes is a major contributor to data governance challenges (Parmiggiani & Grisot, 2020). Additionally, unclear data governance frameworks can erode trust in data, reducing its perceived value and usability. Moreover, Alhassan et al. (2019) demonstrate how missing or ambiguous data governance structures led to distrust, emphasizing the critical role of well-defined data governance frameworks. In their study, the absence of clear data policies necessitated repeated manual data verification, further increasing data work and creating inefficiencies.

The literature also underscores the need to balance human efforts—through guidelines, shared learning, and collaboration—with strict systemic enforcement to ensure effective data governance. While defining data governance activities has been a long-standing focus, research still lacks practical guidelines for adoption and monitoring (Alhassan et al., 2019). Addressing these gaps could benefit scientific research broadly and improve data curation and data acquisition in healthcare, where trust and consistency in data are paramount. Furthermore, data governance has been framed as a collective action problem, particularly in complex organizations where responsibilities and priorities differ (Benfeldt et al., 2020). Given the gaps in the literature on data governance within such complex settings, data governance can be viewed as a mechanism for aligning organizations through shared data governance structures. While data governance is widely recognized for its potential to create value, this value is shaped by how responsibilities are distributed.

Alhassan et al. (2019) identify six essential elements of data governance: (i) cross-functional effort—collaboration and role assignments crucial for governance; (ii) framework—structured support materials, including diagrams and organizational structures; (iii) data as a strategic asset—treating data as a valuable resource with defined goals; (iv) decision rights—determining authority over data pipelines and governance decisions; (v) data policies, standards, and procedures—establishing rules to

support governance objectives; and (vi) compliance monitoring—ensuring adherence to established standards and policies. This highlights the deep interconnection between data governance and the data work needed to make data useful. Effective data governance cannot function without extensive data work, reinforcing the need for structured processes and continuous oversight.

The literature on data governance highlights the importance of sustaining effective data governance over time, as it is critical to long-term success (Ladley, 2019), and one of the pillars of effective data governance is the notion of data quality. Poor data quality is often outlined as the root cause of the most significant challenges in data governance (Janssen et al., 2020). Given that high-quality data is fundamental to rigorous decision-making and research, the extensive efforts required to establish and maintain good data quality as an embedded element of data governance frameworks should be widely acknowledged. While existing research recognizes the importance of data quality, there remains a gap in understanding the data work involved in curating and maintaining high-quality data, a gap which our paper addresses. Poor data quality can lead to significant social and economic consequences, with ripple effects through decision-making and research over time (Wang & Strong, 1996). While organizations are adopting practical strategies to enhance data quality, these efforts often focus primarily on accuracy (Redman, 1998; Steuperaert et al., 2025). However, there are multiple aspects of good data quality. *Firstly*, intrinsic data quality refers to the inherent quality of the data itself. *Secondly*, contextual data quality emphasizes that data quality must align with the specific task or use case. *Thirdly*, representational data quality refers to how well data is structured and presented, and fourth, accessibility data quality underscores the significance of system-related factors in data quality and the notion of making data understandable and usable for its intended purpose. These elements combined illustrate that high-quality data should be inherently reliable, suitable for the given context, well-represented, and easily accessible to users (Mohammed et al., 2025; Steuperaert et al., 2025; Wang & Strong, 1996).

The literature on data quality has been growing in recent years, brought on by the notion that quality data is needed to make quality decisions through machine learning models (Gröger, 2021; Mohammed et al., 2025) as well as generic decision-making, based on quality data. The current stream of literature is focused on data lifecycle concerns and effective data management, much of which is focused on insightful machine learning. However, data quality does not simply happen. Instead, it takes considerable data work, which is largely hidden and invisible. Furthermore, without data work, the data would neither be usable for machine learning purposes nor would it become

insightful for its upstream use. In this paper, we focus on the data work needed to cultivate data governance on the one hand and data quality aspects on the other hand.

### 3 Research approach

Our study is based on data collected through a mixed-methods research approach. The Sleep Revolution is a multidisciplinary initiative focused on advancing sleep research by collecting and preparing data for future studies. The project integrates prospective datasets with up to 20 years of retrospective clinical data sourced from hospitals, research centers, and universities across Europe. Its primary objective is to support ongoing research while establishing a solid foundation for future scientific inquiry related to researching sleep. Researchers can securely access and analyse this data through a high-performance cluster, making the data, embedded within our digital infrastructure, the project's most valuable asset.

#### 3.1 The digital infrastructure

To manage the project's complex and heterogeneous data, we designed a lean database framework intended to reduce friction in data work across disciplinary boundaries (see Figure 1). The framework organizes over 2,600 variables from diverse datasets into seven unified tables, each linked to individual participants and studies via a shared primary key. This relational model supports a broader digital infrastructure used by 39 research centers, with most being in Europe, integrating a range of data types, including sensor data, mobile app logs, clinical records, and survey inputs (see more details in Sveinbjarnarson et al., 2023, 2024).

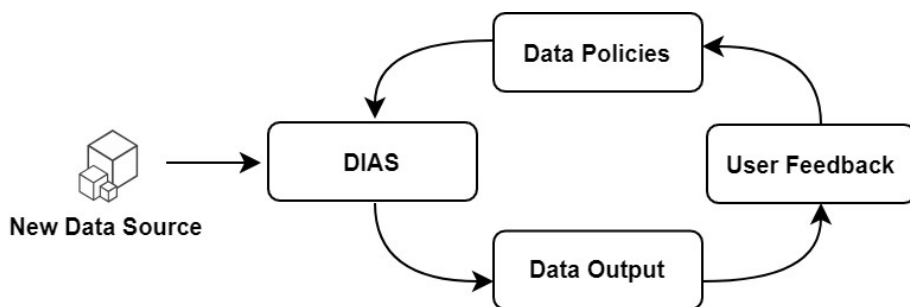


Figure 1. Illustrates the iterative process used to create data policies and rules in the project.

The design is modular and schema-light, centering on four core entities: (i) Form, which defines the table template; (ii) Entry, representing individual data fields or parameters; (iii) FormResult, corresponding to rows of submitted data; and (iv) EntryResult, storing cell-level values. This structure allows new data sources to be added without altering the table architecture, supporting extensibility while maintaining structural consistency. Metadata fields further support evolving data needs, enabling documentation and traceability without additional schema complexity.

A key component of this digital infrastructure is a built-in validation system that enforces standards. This validation system evolved into the Data Integrity Assurance System (DIAS), developed iteratively over four years. DIAS was shaped by user feedback and co-design sessions, leading to a system capable of enforcing format and logic rules while remaining usable by non-technical contributors. The goal was to offload quality enforcement from manual review to automated validation without requiring users to write scripts or understand the database internals.

As DIAS matured, it became central to the infrastructure's approach to data governance. Its rules and policies were refined through continuous use, addressing issues such as inconsistent identifiers, formatting errors, and structural mismatches. Improvements focused on enhancing findability, usability, and data consistency under resource constraints. This chapter outlines the architecture and development of this infrastructure as a basis for evaluating how users interact with automated data validation systems and whether such systems can support independent standard application and improved data quality in practice.

## 3.2 Study design, data collection, and data analysis

This paper relies on the digital infrastructure, which we have designed and developed for the past four years, as well as a mixed-methods empirical data collection, combining a feasibility test, survey, and semi-structured interviews to evaluate the data work behind applying data standards independently through automation and to investigate how user behavior, preferences, and data practices impact data quality. The feasibility test focused on whether the DIAS validation system could be used without assistance, relying only on a brief introduction and documentation. The broader aim of our empirical data gathering was to assess whether such validation systems could shift responsibility for standardization and validation from technical personnel to the individuals who collected and understood the data; broadening the data work.

The participants were engaged in a structured process lasting between 65 and 120 minutes (see Figure 2). All 13 participants who partook in our data gathering had prior

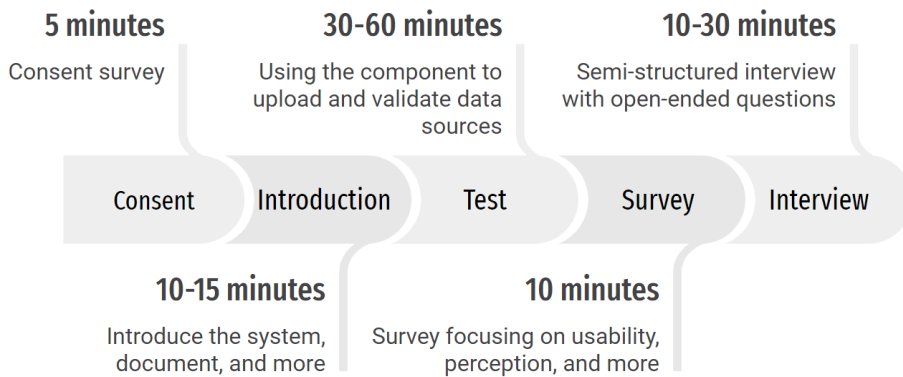


Figure 2. Participant engagement and task sequence.

experience working with healthcare data. At the beginning of the session, participants completed a consent form explaining the study's purpose, the data being collected, their right to withdraw at any time, and how their information would be handled. Following this, the first author provided an oral introduction to the system, including an overview of the files provided, the task ahead, and how to interpret the documentation. Each participant received a ZIP file containing example input files, a small CSV dataset (12 rows, 8 columns), documentation, and a README file. After the introduction, they were asked to convert the data into six structured CSV files and upload them to the validation system.

Each file corresponded to a specific table in the relational database; participants, studies, study participation, forms, entries, and form results. The complexity of the tasks varied depending on the structure and formatting requirements of each file. More complex files, such as form results, required column-level metadata and stricter formatting, while others involved simpler mappings. The goal was to simulate realistic scenarios in which users prepare structured data independently and rely on documentation and feedback from the validation system to complete the task. The feasibility test allowed us to observe how participants interacted with the system and highlighted challenges in understanding and applying standards.

The participants uploaded their files to a minimalist website connected to the database. After each upload, the validation system provided immediate feedback indicating whether the submission met predefined standards or contained any violations (see example output in Appendix A). In remote sessions, participants shared their screens, allowing the researcher to observe how they engaged with documentation and responded to feedback. The first author answered clarifying questions about data content during the introduction of the system, but after participants began the task, further assistance was avoided unless participants became completely stuck, which did not occur. In such

cases, they were simply reminded to refer to the example files and documentation provided.

During the data collection, observational data were collected on task completion rates, time spent, standard violations encountered, participants' use of documentation, and the files uploaded by participants were stored by the system for further review. Following the feasibility test, participants completed a survey capturing their research backgrounds, challenges encountered in data work, time investment in data work, and perspectives on data quality. To deepen our understanding, a semi-structured interview with 17 questions was conducted, exploring participants' experiences with the validation system, its design, and broader views on data governance and data handling.

The qualitative data were analyzed using an abductive thematic approach. We engaged in an interplay between real-world observations through our interview transcripts and insights from existing literature on data work more broadly (Gregory and Muntermann, 2011), going between the two during our analysis. The thematic analysis was conducted in three phases, through iterative coding of the interview transcripts, starting with open descriptive tags. *The first phase* was conducted by the first author and involved primary coding, assigning initial label segments with descriptive and conceptual tags. This phase was open-ended and exploratory, focused on capturing the essence of the data yet trying to organize and interpret it deeply (Tracy, 2024). *The second phase* was also conducted by the first author and involved secondary coding, revisiting the codes, and beginning to group them, refining them, and identifying the relationship between the codes. These were grouped into broader categories to identify patterns such as difficulties in interpreting data, engagement with documentation, and strategies for addressing validation feedback. *The third phase* was conducted in collaboration between the first and the last author of this paper and included the theme creation. This phase involved axial coding as well as interpretation and abstraction, which ultimately led to the three themes presented in the findings in this paper. The three key elements presented in the findings were drawn from both the frequency of these categories and their relevance to the study's aims, including participants' perceptions of the system and responses to questions related to data work, derived through the third phase. The quantitative survey data were also coded and categorized. Full details of the survey and interview instruments are provided in Appendix A.

### 3.3 Recruitment and ethical considerations

The participants were recruited through convenience sampling via professional networks and referrals, with a focus on individuals actively engaged in data work. All had

prior experience working with healthcare data and came from diverse fields, including sports science, neuroscience, information systems, and machine learning. Eligibility criteria required prior data experience, English proficiency, no involvement in the development of the validation system, and access to spreadsheet software.

The interviews were audio recorded, transcribed verbatim, and lightly edited for clarity. Non-English interviews (conducted in Icelandic) were translated into English before analysis. The collected data included survey responses, uploaded task files, interview transcripts, and observational notes. The survey data were cleaned for consistency prior to analysis. Although multiple ethical approvals were obtained for the broader project, this specific study did not require ethical approval under Icelandic regulations, namely Act No. 90/2018 on Data Protection and the Processing of Personal Data, as no personal or sensitive data were collected. All participants, who were researchers, signed informed consent before participating. They were informed that their anonymized data would be used for research purposes and that they could withdraw their consent at any time, either by contacting the research team or by revisiting the consent form and changing their response from 'yes' to 'no'. This study is part of the Sleep Revolution project and was funded by the European Union's Horizon 2020 research and innovation programme under grant agreement no. 965417.

## 4 Results

To explore the complexities of data work and its implications for data governance, we identified six key elements essential for curating healthcare data and advancing healthcare research. *First*, data work is prone to frequent errors, leading to a high mistake rate. *Second*, researchers often rely on assumptions when handling data, even when unfamiliar with its context. These first two elements of data work are illustrated in the section called *the data work needed for reaching intrinsic data quality in data governance*. Intrinsic data quality refers to the inherent quality of the data itself, and our paper shows that there is data work needed to unfold intrinsic data quality and that data work involves gaining a deep understanding of data needs and trying to lessen the likelihood of errors and assumptions.

*Third*, there is a widespread disregard for standards, regulations, and documentation, demonstrating that human-driven data work remains essential in balancing systemic measures and human effort despite structured data governance. *Fourth*, perceptions of data quality vary significantly—what one considers crucial, such as metadata or consistent formatting, may be dismissed as unimportant by another. These two elements of data work are illustrated in the section called *the data work needed for reaching contextual*

*data quality in data governance.* Contextual data quality emphasizes that data quality must align with the specific task or use case, and our paper shows that there is data work needed to align the human and systemic measures.

*Fifth*, data users often lack awareness of the broader needs of others when collecting and processing data for shared use. *Sixth*, many researchers have never engaged with data standards, data principles, or data governance, resulting in significant, often unseen data work later in the research process. These two elements are illustrated in the section called *the data work needed for reaching representational and accessibility data quality in data governance.* Representational data quality refers to how well data is structured and presented, and accessibility data quality underscores the significance of system-related factors in data quality. Our paper shows the importance of doing so through digital infrastructures and cultivating excellence in data work early on, as that, in turn, reduces later workload.

## 4.1 The data work needed for reaching intrinsic data quality in data governance

During the feasibility evaluation, we analysed the type of data work needed within the digital infrastructure for reaching intrinsic data quality and how the data could be checked for its feasibility for the purpose at hand. Our findings show that a deep understanding of data needs is the essence of data work, and that, in turn, elevates data curation. Our participants were tasked with splitting a provided spreadsheet into six new files, adding metadata columns to three of them to examine their approach to data work. They were given example input sheets for each file type, a short description file, and detailed documentation, which they were encouraged to use. The external tools participants chose varied, including Excel, Google Sheets, LibreOffice Calc, and text editors. While some modified the example files to incorporate the new data, others created entirely new files—both approaches reflecting different forms of data work with potential ripple effects on data governance down the line.

All participants successfully uploaded the six required files as part of our study. However, none completed the task without encountering at least one error flagged by the validation system within the digital infrastructure. Although the provided data was free of initial errors or formatting issues, participants introduced new errors, triggering validation checks that identified problems such as incorrect formatting, missing or additional columns, duplicate entries, incorrect file types, and structural inconsistencies. These errors highlight critical aspects of data governance that must be addressed to ensure sustainable data quality.

Common mistakes included spelling errors, incorrect date formatting, and the accidental deletion of rows perceived as duplicates. Such errors, if not caught early, could necessitate additional data cleaning efforts later, increasing the data work required for future research. Some errors, such as misinterpreted metadata or mislabelled columns, were not flagged by the validation system but were identified through direct observations. These observations underscore the importance of early intervention in preventing faulty data from entering research databases, ultimately safeguarding the integrity of data used for healthcare analysis and scientific discovery.

Participant behaviour exhibited several distinct trends. Initially, many displayed low confidence, frequently seeking clarification rather than engaging with the documentation or system feedback. Over time, they became more efficient and confident in using the digital infrastructure, though frustration was common when uploads were rejected. Notably, many participants approached the task as a challenge to “get the system to accept their input” rather than an exercise in meaningful data work. This mindset led them to prioritize satisfying validation requirements over ensuring that the data was properly structured and meaningful for future use.

Participants expressed diverse preferences on how structured and raw data should be shared. Some favoured storing data in a database with query capabilities, while others preferred simpler formats like folder structures or Excel files. These varying perspectives highlight the lack of a universal approach to data storage and underscore the need for adaptable digital infrastructures that accommodate different workflows.

Overall, the results reveal a pattern of minimal engagement with documentation, a reliance on assumptions, and a reactive rather than proactive approach to data preparation. The high frequency and diversity of errors detected by the digital infrastructure emphasize the critical role of robust validation processes in sustainable data governance. Additionally, these findings highlight the need for improved user guidance and education on the significance of preparing data for long-term reuse. The participants’ engagement with documentation revealed another critical challenge. Although documentation was provided and participants were encouraged to use it, only a minority actively relied on it during their tasks. Participant 13 admitted, “I think they are important, but no one reads them.” Even those who found documentation helpful engaged with it sparingly. Participant 1 noted, “I didn’t work much with the documentation, but I found it helpful for specific tasks, like how to write institution names.” Similarly, Participant 10 stated, “In general, the documentation is essential, but I did not use it except for one thing I was unsure of.” These responses suggest that relying solely on documentation to ensure proper data handling is insufficient. Instead, cultivating excellence in data work requires a shift in practice—one that emphasizes ongoing discussions, training, and

embedding data quality practices into daily research activities. Based on that, we offer two lessons learned. Data work is prone to frequent errors, leading to a high mistake rate, and furthermore, researchers often rely on assumptions when handling data, even when unfamiliar with its context.

## 4.2 The data work needed for reaching contextual data quality in data governance

Despite the high number of errors detected by the digital infrastructure during testing, survey results indicate that participants generally found the system and the process of creating the required data files to be relatively straightforward, even as first-time users. Most participants rated most of the system functions as ‘very easy’ to relate to and work with or ‘somewhat easy’, with only a few selecting ‘neutral’ in terms of difficulty when working within the data governance structure. Three participants rated the final task as ‘somewhat difficult’, but no one considered any task ‘very difficult.’ Overall, nearly all participants agreed or strongly agreed that the digital infrastructure and its embedded data governance mechanisms were beneficial for research projects, particularly within healthcare, with only one participant expressing a neutral stance.

Interestingly, our survey findings revealed significant disagreement among participants regarding what constitutes ‘good’ data quality. While some emphasized the importance of consistent data formats, unique research identifiers, and the elimination of duplicates, others considered these aspects less critical. This variability underscores the diverse priorities and perspectives researchers bring to data governance, highlighting the need for digital infrastructure that can accommodate differing interpretations of data quality. However, one area of strong consensus emerged: the importance of identifying which dataset a specific data source belongs to. Most participants considered this essential, particularly in healthcare, where accurately linking patient data to symptoms and associated records is crucial for both research and clinical decision-making.

Participants also highlighted challenges in working with data they had not collected themselves, often revealing underlying trust issues. Many struggled to understand parameters or variables, especially when labelled with abbreviations or general terms they had not defined. These ambiguities directly contributed to distrust in the data. The absence or inaccessibility of metadata further compounded these issues, making interpretation and reuse difficult and reinforcing scepticism about data reliability. Additionally, merging datasets proved challenging due to missing identifiers and inconsistent formats, illustrating the complexities researchers face in ensuring data compatibility as a crucial aspect of data work.

When discussing broader challenges in data handling, participants pointed to several key factors: educational gaps, human error, and technological limitations. Participant 7 identified multiple issues, stating, “Human errors, not understanding what is important, how to format, not complete... poor documentation, time-consuming.” Education was also seen as a critical factor in mitigating these issues. Participant 3 argued that “I think with data processing and programming as a basic education from primary school and upwards, this would be much less of a problem.” Participant 5 further noted, “The learning process is different for different people... [data is] not easy to see where the errors surface.” Participants also expressed concerns about inconsistencies in data collection and reliance on temporary fixes rather than long-term solutions. Participant 8 explained, “Temporary fixes instead of preventing errors [...] We are limited by older systems or solutions that have been used for a long time.”

Furthermore, several participants acknowledged the tendency to prioritize immediate efficiency over long-term usability. Participant 13 observed, “When people start working on something, they don’t think enough about the future [...] structure, and all their work so that it could be used in the future.” This sentiment reflects a key challenge in data governance—balancing short-term ease of data handling with sustainable practices that ensure long-term data quality. The lack of adherence to data principles, inconsistencies in formats, and poor conversion tools further complicate data governance. As Participant 11 succinctly put it, “Not everyone follows the principles. All of those different formats are difficult, and there are formatting issues all over.” These findings emphasize the need for structured but adaptable data governance models that not only enforce quality standards but also integrate best practices for data work in a way that researchers find practical and useful.

Our participants reported spending considerable time on data work tasks such as data conversion, data merging, and data cleaning—processes that should be more streamlined with standardized formats. These inefficiencies in current data governance practices highlight the need for digital infrastructure like the one presented here, which aims to ensure that data is clean, validated, and ready for use from the outset. However, our findings also reveal tensions and reduced trust when researchers work with data, they did not collect themselves, reinforcing the notion that data work is inherently complex and deeply human. Based on these findings, we offer two additional lessons learned in addition to those already posed in the first subsection of the findings. There is a widespread disregard for standards, regulations, and documentation, demonstrating that human-driven data work remains essential for balancing systemic measures and human effort despite structured data governance. Moreover, perceptions of data quality

vary significantly; what one considers crucial, such as metadata or consistency in formatting, may be dismissed as unimportant by others.

### 4.3 The data work needed for reaching representational and accessibility data quality in data governance

In our interviews, we examined participants' training, readiness, experience in handling data, familiarity with data quality principles, and perspectives on why data is often problematic. Through our analysis, it was clear that data work remains undervalued and is rarely treated as a formal skill. Many participants received little to no formal training in data handling or data quality. Instead, they relied on self-taught methods or guidance from colleagues, learning 'on the fly' rather than through structured education. This lack of formalized training suggests that data work remains largely invisible, assumed rather than taught, and often overlooked as a crucial component of research.

Participant 8 noted: "No, everything [was] self-taught," and Participant 10 echoed: "No, only brief courses on handling data. Maybe something brief, but I cannot remember." Similarly, Participant 4 stated, "Not really, everything I've learned has been through work." Even those with formal training described a limited scope, typically focusing on technical aspects like database design or data cleaning. Participant 5 shared: "I did my master's in data science [...] We had courses on data quality and cleaning," while Participant 6 added: "Yes, I have had many courses on structuring and creating data in databases and other structures."

The absence of structured education in data governance points to a systemic challenge that hinders collective action and collaboration in data work. Researchers approach data inconsistently without a shared foundation of best practices, leading to errors and inefficiencies. Our findings suggest that cultivating excellence in data work requires more than individual expertise—it necessitates greater visibility, institutional recognition, and structured training to ensure data quality is not left to assumption or chance.

We also explored the extent to which participants engaged with data principles or formal standards in their data work. Most had little exposure to established data standards, and those who did often relied on ad-hoc practices or conventions specific to individual research projects rather than universally recognized frameworks. Participant 2 stated, "None I can remember, no guidelines or standards in my current or previous work." Similarly, Participant 7 noted, "No, no documentation, but maybe somewhat through adapting others' work." This absence of structured principles suggests that data

management often lacks uniformity, making it difficult to share and reuse data across research teams.

Some participants attempted to create their own standards to maintain consistency. For instance, Participant 12 shared, “Yes, but I made it myself. I organized the data we store and the [sleep] data format.” However, the implementation was often incomplete even among those who followed the guidelines. As Participant 4 pointed out, “Yes, some, but way too few. [We use] naming standards for sleep scoring, for example.” These findings indicate that without widely accepted data governance structures, researchers develop their own fragmented approaches, potentially reinforcing inconsistencies and increasing data work over time.

When asked about implementing the digital infrastructure for validation in their own work, participants largely supported its use, particularly for improving data consistency and strengthening data governance. Participant 1 remarked, “Yes. The data is so much. Especially good for participants in multiple studies.” Similarly, Participant 9 stated, “Yes, that would be highly beneficial as long as you have access to the source data for transparency and trust.” Others emphasized its role in unifying data across research projects. As Participant 6 noted, “I think we need to have it to unify all of the data.” Our findings also show that although it might make sense to implement data governance in general, pressured everyday practice often leads to opting for minimal viable effort due to resource scarcity. It can, however, shift the burden onto others, creating a cascade effect that leads to more work down the line.

Furthermore, some participants raised concerns about flexibility, suggesting that strict validation rules might hinder creativity in certain research contexts. Participant 3 elaborated, “Having a strict format that your data has been in could become an unnecessary hurdle [...] It really depends on what kind of experiment it is.” These findings highlight the balance between enforcing data standards and allowing adaptability in data governance. Based on that, we offer two more lessons learned. Data users often lack awareness of the broader needs of others when collecting and processing data for shared use. Furthermore, many researchers have never engaged with data standards, data principles, or data governance, leading to extensive and often invisible data work later in the research process.

## 5 Discussion

Knorr-Cetina’s (1997) influential research on knowledge work suggests that modern work increasingly mirrors the characteristics of scientific research. This perspective aligns with Barley and Bechky’s (1994) notion of the “backrooms of science,” which

highlights the often-overlooked labour of laboratory technicians in preparing, managing, and curating scientific objects and research data. Building on this foundation, Parmiggiani et al. (2022) emphasize the significant yet hidden efforts required to make data ready for analysis and knowledge curation. Expanding on these insights, our paper explores the complex and nuanced data work undertaken by researchers in healthcare settings—work that is critical for the design, development, and sustainability of data-driven digital infrastructure and fundamental to advancing healthcare decision-making and research through quality data. Combined, our findings highlight several insights that shed light on data work for mediating data quality and data governance in a data-intensive world. *First*, data work is prone to frequent errors, leading to a high mistake rate. *Second*, researchers often rely on assumptions when handling data, even when unfamiliar with its context. *Third*, there is a widespread disregard for standards, regulations, and documentation, demonstrating that human-driven data work remains essential in balancing systemic measures and human effort despite structured data governance. *Fourth*, perceptions of data quality vary significantly—what one considers crucial, such as metadata or consistent formatting, may be dismissed as unimportant by another. *Fifth*, data users often lack awareness of the broader needs of others when collecting and processing data for shared use. *Sixth*, many researchers have never engaged with data standards, data principles, or data governance, leading to extensive and often invisible data work later in the research process.

## 5.1 Data quality is fostered through a symbiosis between human effort & systemic measures

Data governance has emerged as a key pillar for ensuring data quality in both organizations and research (Benfeldt et al., 2020). However, as our findings show, effective governance requires continuous oversight and dedicated teams managing data throughout its lifecycle. This includes ongoing data work to ensure availability, inform stakeholders, and address challenges. Despite its importance, research grants rarely fund dedicated data governance, focusing instead on outputs like publications. As a result, data work becomes frequently sidelined—even though high-quality research depends on high-quality data (Khong et al., 2023).

The growing literature on data quality reflects its role in enabling sound decisions through artificial intelligence and machine learning (Gröger, 2021; Mohammed et al., 2025). Yet, data governance and quality efforts require personnel and resources that many research projects lack. However, this paper is a call for action; data work in healthcare research must be recognized, prioritized, and resourced appropriately, as

data work is pivotal for data quality. Our findings highlight the deep interconnection between data governance and data work. Beyond technical aims, investing in human expertise is essential for leveraging data's full potential in healthcare innovation. Numerous data principles and governance frameworks exist (Janssen et al., 2020), but adherence is often merely partial. While this may reduce their immediate data work, our findings show that it instead frequently shifts the burden onto others, creating a cascade effect. Shortcuts taken to minimize effort at one stage can lead to significant additional data work for others later, which in turn may ultimately undermine data quality and data governance effectiveness.

Our feasibility evaluation showed that—despite being encouraged—only two participants fully utilized the provided documentation, leading to errors directly linked to their approach. This underscores the challenge of ensuring compliance with data governance models, even when their benefits are well understood (Benfeldt et al., 2020). Our findings emphasize the value of early-stage data work and meaningful human engagement, which can greatly enhance downstream scientific inquiry. They also reveal the vast scope and critical importance of often-overlooked data work in healthcare research. Neglecting data at the point of entry can lead to misinformation, flawed analyses, and unreliable findings, particularly dangerous in healthcare (Adams et al., 2023). Ensuring quality data requires effort in collection, curation, and validation, highlighting the need to acknowledge and invest in the labour behind data-driven healthcare research. Poor data quality can have serious social and economic consequences, rippling through decision-making and research (Wang & Strong, 1996).

Our findings revealed a high error rate. No participant uploaded all files without at least one validation error. Common mistakes included accidental row deletions and incorrect date entries. Despite built-in data governance mechanisms, adhering to guidelines remained difficult. This reinforces the need for both human effort and systemic measures. The errors we observed show that unless proactively prevented, data errors will inevitably occur. This highlights a critical trade-off between relying solely on systemic controls and investing in human awareness of data quality. Meaningful progress requires both strong technological safeguards and a deeper understanding of the importance of data work. As Jarvenpaa & Essén (2023) argue, data governance must bridge technological and human dimensions. Our paper supports this, showing that sustainable data quality demands both systemic measures and human effort.

A closer analysis of the errors reveals a key insight: participants were tasked with uploading a small dataset, only eight columns and 13 rows. In contrast, our digital infrastructure integrates 63 data sources, covering 2,632 parameters from around 2,000 participants, resulting in millions of data entries. Unlike controlled test data or struc-

tured financial data, real-world healthcare data is full of inconsistencies—spelling errors, duplications, formatting issues—highlighting the complexity and variability of such data. This increases the likelihood of validation errors and underscores the ongoing need for robust data work. Healthcare data, inherently tied to human variability, is neither static nor simplistic.

These findings underscore that while data governance provides essential guidelines, embedded validation systems are vital for reducing errors (Alhassan et al., 2019). More broadly, data governance requires a careful symbiosis between human effort and systemic measures to ensure data quality. These findings highlight the importance of further research into trust in digital technology and its data (Li et al., 2008; Müller et al., 2024), particularly in the context of data governance. Establishing trust is crucial for fostering user adoption and ensuring that data work—the foundation of effective data governance—becomes an integral and well-supported part of research workflows.

## 5.2 A key element of data governance is making data work visible

Despite all participants being healthcare researchers, their views on data quality, storage, and challenges varied significantly. Some emphasized eliminating duplicates, consistent formatting, and thorough documentation, while others downplayed these aspects. This divergence likely stems from their different backgrounds and environments. For instance, researchers with database experience found duplicate removal easy, whereas others lacked the knowledge to handle such issues. These findings reflect the complexities of multidisciplinary research and highlight the importance of effective data governance and collaboration across disciplines. While literature stresses data governance, our study shows its practical adoption is challenging and demands both, as stated earlier, systemic measures and a cohesive shared understanding of data principles.

Collaborative data work is hindered by differing priorities and expertise, leading to inconsistent attention to data quality. This was evident in our study, as only two participants who managed metadata effectively also stressed documentation as vital in the interviews. Their views directly shaped their data practices. While all participants agreed on data's centrality in today's research landscape, the data work to make it usable often remains invisible and undervalued (Abbasi et al., 2016; Bossen et al., 2019). To address this, we advocate for open discussions to build shared understanding and actively promote that data work is made visible in healthcare research.

Although most researchers could identify causes of poor data quality, over half had no formal training in data standards or principles. Instead, they relied on intuition and tri-

al-and-error. Minimal standards—such as naming conventions or storage locations—were typically group-specific. This contradiction underscores the invisibility of data work and the need to prioritize training and support, and the importance of making data work increasingly visible.

Quality data requires thoughtful data work, supported by validation systems that bridge human effort and systemic measures. Our findings point to a lack of mutual understanding between disciplines, further complicated in collaborative, multidisciplinary research. This aligns with the seminal work of Schmidt and Bannon (1992) on articulation work, which emphasizes the importance of coordinating and communicating invisible labour. Decisions about data storage, standards, and governance are always made by someone. However, they often fail to account for the needs of all stakeholders (Alhassan et al., 2019). We argue that such decisions should not remain hidden in the background. Instead, they must be visible and recognized as essential elements of data work in healthcare research.

The literature has shown that there are multiple aspects of what can be seen as good data quality, and our paper illustrates the data work connected to each of those, contributing to the literature within information systems on data governance in general and data quality in particular. *Firstly*, intrinsic data quality refers to the inherent quality of the data itself, and our paper shows that data work is needed to reach intrinsic data quality by gaining a deep understanding of data needs, which is the essence of data work, and that, in turn, elevates data curation. We show that data work is prone to frequent errors, leading to a high mistake rate, and show that researchers often rely on assumptions when handling data, even when unfamiliar with its context. *Secondly*, contextual data quality emphasizes that data quality must align with the specific task or use case, and our paper shows that data work is needed to address trust issues, which in turn can prevent additional data work downstream. Data users often lack awareness of the broader needs of others when collecting and processing data for shared use. Our paper shows that there is a widespread disregard for standards, regulations, and documentation, demonstrating that human-driven data work remains essential in balancing systemic measures and human effort despite structured data governance. We also show that perceptions of data quality vary significantly—what one considers crucial, such as metadata or consistent formatting, may be dismissed as unimportant by another. *Thirdly*, representational data quality as well as accessibility data quality underscore the significance of system-related factors in data quality, and our paper shows the importance of doing so through digital infrastructures and cultivating excellence in data work early on, as that, in turn, reduces later workload. We show that many researchers have never engaged with data standards, data principles, or data governance, leading to extensive

and often invisible data work later in the research process and a lack of contextual data quality understanding, which could be cultivated through making that element of the data work visible. Ergo, high-quality data should be inherently reliable, suitable for the given context, well-represented, and easily accessible to users, and high-quality data work is needed to reach each of these. *Lastly*, we argue that data work must be made visible and recognized as an essential component of scientific research. Furthermore, decisions regarding data quality standards and data governance frameworks should not be confined to the backrooms of science (Barley & Bechky, 1994) or data science (Parmiggiani et al., 2022). Instead, data work should be treated as visible and integral elements of research, shaping how healthcare data is thoughtfully collected, processed, maintained, and used.

### 5.3 Limitations and future work

Despite the valuable insights from our study, several limitations must be acknowledged. *First*, our participant pool included only researchers working with research data; ergo, our findings may not generalize to other disciplines or settings. *Second*, although our sample size was small (13 participants), we mitigated this by providing rich, detailed descriptions of observed data work, supported by qualitative interview insights. Future research could develop scalable data governance frameworks for multidisciplinary research, particularly in healthcare, with an emphasis on making essential data work visible.

## 6 Conclusion

In this paper, we examine the data work required to establish data-driven digital infrastructure in healthcare and researchers' efforts to prepare healthcare data for its use. Our findings highlight several insights that shed light on data work for mediating data quality and data governance in a data-intensive world. Our paper illustrates the data work needed for reaching intrinsic data quality in data governance, (ii) the data work needed for reaching contextual data quality in data governance, and (iii) the data work needed for reaching representational and accessibility data quality in data governance.

The paper explores various dimensions of data quality—intrinsic, contextual, representational, and accessibility dimensions—and the essential role of data work in achieving each. We highlight the ways in which data work can be error-prone, assumption-driven, and affected by disregard for standards and regulations. Our paper emphasizes the need for understanding data context, fostering trust, and early engagement

with digital infrastructure to reduce future data work. We also argue that data work, though frequently invisible, is vital for scientific research and should be recognized as such. Ultimately, decisions about data quality standards and data governance frameworks must be recognized as central to research practice, shaping how healthcare data is collected, curated, and used, rather than being relegated to invisible, backroom work.

## Bibliography

- Abbasi, A., Sarker, S., & Chiang, R. (2016). Big Data Research in Information Systems: Toward an Inclusive Research Agenda. *Journal of the Association for Information Systems*, 17(2). <https://doi.org/10.17705/1jais.00423>
- Adams, Z., Osman, M., Bechlivanidis, C., & Meder, B. (2023). (Why) Is Misinformation a Problem? *Perspectives on Psychological Science*, 18(6), 1436-1463. <https://doi.org/10.1177/17456916221141344>
- Alhassan, I., Sammon, D., & Daly, M. (2016). Data governance activities: An analysis of the literature. *Journal of Decision Systems*, 25(1), 64-75. <https://doi.org/10.1080/12460125.2016.1187397>
- Alhassan, I., Sammon, D., & Daly, M. (2019). Critical Success Factors for Data Governance: A Theory Building Approach. *Information Systems Management*, 36(2), 98-110. <https://doi.org/10.1080/10580530.2019.1589670>
- Arnardottir, E. S., Islind, A. S., & Óskarsdóttir, M. (2021). The Future of Sleep Measurements: A Review and Perspective. *Sleep Medicine Clinics*, 16(3), 447-464. <https://doi.org/10.1016/j.jsmc.2021.05.004>
- Barley, S. R., & Bechky, B. A. (1994). In the Backrooms of Science: The Work of Technicians in Science Labs. *Work and Occupations*, 21(1), 85-126. <https://doi.org/10.1177/0730888494021001004>
- Benfeldt, O., Persson, J. S., & Madsen, S. (2020). Data Governance as a Collective Action Problem. *Information Systems Frontiers*, 22(2), 299-313. <https://doi.org/10.1007/s10796-019-09923-z>

- Bertelsen, P. S., Bossen, C., Knudsen, C., & Pedersen, A. M. (2024). Data work and practices in healthcare: A scoping review. *International Journal of Medical Informatics*, 184, 105348. <https://doi.org/10.1016/j.ijmedinf.2024.105348>
- Bossen, C., Chen, Y., & Pine, K. H. (2019). The emergence of new data work occupations in healthcare: The case of medical scribes. *International Journal of Medical Informatics*, 123, 76-83. <https://doi.org/10.1016/j.ijmedinf.2019.01.001>
- Cruz, T. M. (2023). Data politics on the move: Intimate work from the inside of a data-driven health system. *Information, Communication & Society*, 26(3), 496-511. <https://doi.org/10.1080/1369118X.2021.1954972>
- Davenport, T. H., & Patil, D. (2012). *Data scientist*. Harvard Business Review, 90(5), 70-76.
- Fiske, A., Prainsack, B., & Buyx, A. (2019). Data Work: Meaning-Making in the Era of Data-Rich Medicine. *Journal of Medical Internet Research*, 21(7), e11672. <https://doi.org/10.2196/11672>
- Galliers, R. D., Newell, S., Shanks, G., & Topi, H. (2017). Datification and its human, organizational and societal effects. *The Journal of Strategic Information Systems*, 26(3), 185-190. <https://doi.org/10.1016/j.jsis.2017.08.002>
- Gregory, K., Groth, P., Cousijn, H., Scharnhorst, A., & Wyatt, S. (2019). Searching Data: A Review of Observational Data Retrieval Practices in Selected Disciplines. *Journal of the Association for Information Science and Technology*, 70(5), 419-432. <https://doi.org/10.1002/asi.24165>
- Gregory, R., & Muntermann, J. (2011). Theorizing in design science research: inductive versus deductive approaches. In 32nd *International Conference of Information Systems*.
- Gröger, C. (2021). There is no AI without data. *Communications of the ACM*, 64(11), 98-108. <https://doi.org/10.1145/3448247>

- Hodge, B., Richert, E., Schmitz, L., Islind, A. S., Leppänen, T., McNicholas, W. T., ... & Pires, G. N. (2025). Behavioral interventions as treatment for insomnia-A scoping review of randomized controlled trials. *Sleep Medicine*, 108676.
- Islind, A., Vallo Hult, H., Johansson, V., Angenete, E., & Gellerstedt, M. (2021). Invisible Work Meets Visible Work: Infrastructuring from the Perspective of Patients and Healthcare Professionals. *Proceedings of the 54th Hawaii International Conference on System Sciences*. <https://doi.org/10.24251/HICSS.2021.431>
- Janssen, M., Brous, P., Estevez, E., Barbosa, L. S., & Janowski, T. (2020). Data governance: Organizing data for trustworthy Artificial Intelligence. *Government Information Quarterly*, 37(3), 101493. <https://doi.org/10.1016/j.giq.2020.101493>
- Janssen, M., Wallenburg, I., & de Bont, A. (2016). Carving Out a Place for New Health Care Occupations: An Ethnographic Study into Job Crafting. In H. Albach, H. Meffert, A. Pinkwart, R. Reichwald, & W. von Eiff (Eds.), *Boundaryless Hospital: Rethink and Redefine Health Care Management* (pp. 119-141). Springer. [https://doi.org/10.1007/978-3-662-49012-9\\_7](https://doi.org/10.1007/978-3-662-49012-9_7)
- Jarvenpaa, S. L., & Essén, A. (2023). Data sustainability: Data governance in data infrastructures across technological and human generations. *Information and Organization*, 33(1), 100449. <https://doi.org/10.1016/j.infoandorg.2023.100449>
- Jones, M. (2019). What we talk about when we talk about (big) data. *The Journal of Strategic Information Systems*, 28(1), 3-16. <https://doi.org/10.1016/j.jsis.2018.10.005>
- Khong, I., Yusuf, N. A., Nuriman, A., & Yadila, A. B. (2023). Exploring the Impact of Data Quality on Decision-Making Processes in Information Intensive Organizations. *APTISI Transactions on Management*, 7(3), Article 3. <https://doi.org/10.33050/atm.v7i3.2138>
- Knorr-Cetina, K. (1997). Sociality with objects: Social relations in postsocial knowledge societies. *Theory, culture & society*, 14(4), 1-30.

- Ladley, J. (2019). *Data Governance: How to Design, Deploy, and Sustain an Effective Data Governance Program*. Academic Press.
- Li, X., Hess, T. J., & Valacich, J. S. (2008). Why do we trust new technology? A study of initial trust formation with organizational information systems. *The Journal of Strategic Information Systems*, 17(1), 39-71. <https://doi.org/10.1016/j.jsis.2008.01.001>
- Mohammed, S., Budach, L., Feuerpfeil, M., Ihde, N., Nathansen, A., Noack, N., ... & Harmouch, H. (2025). The effects of data quality on machine learning performance on tabular data. *Information Systems*, 132, 102549. <https://doi.org/10.1016/j.is.2025.102549>
- Müller, L. S., Nohe, C., Reiners, S., Becker, J., & Hertel, G. (2024). Adopting information systems at work: A longitudinal examination of trust dynamics, antecedents, and outcomes. *Behaviour & Information Technology*, 43(6), 1096-1128. <https://doi.org/10.1080/0144929X.2023.2196598>
- Parmiggiani, E., & Grisot, M. (2020). Data Curation as Governance Practice. *Scandinavian Journal of Information Systems*, 32(1). <https://aisel.aisnet.org/sjis/vol32/iss1/1>
- Parmiggiani, E., Østerlie, T., & Almklov, P. (2022). In the Backrooms of Data Science. *Journal of the Association for Information Systems*, 23, 139-164. <https://doi.org/10.17705/1jais.00718>
- Petzold, B., Roggendorf, M., Rowshankish, K., & Sporleder, C. (2020). *Designing data governance that delivers value*. McKinsey & Company.
- Redman, T. C. (1998). The impact of poor data quality on the typical enterprise. *Communications of the ACM*, 41(2), 79-82.
- Schmidt, K., & Bannon, L. (1992). Taking CSCW seriously. *Computer Supported Cooperative Work (CSCW)*, 1(1), 7-40. <https://doi.org/10.1007/BF00752449>
- Schmitz, L., Richert, E., Jóhannsdóttir, K. R., Arnardóttir, E. S., & Islind, A. S. (2025). VALID Care Pathways: A Framework for Meaningful Digital Platform Design. *Communications of the Association for Information Systems*, 56(1), 45.

- Su, Z., He, L., Jariwala, S. P., Zheng, K., & Chen, Y. (2022). “What is Your Envisioned Future?”: Toward Human-AI Enrichment in Data Work of Asthma Care. *Proc. ACM Human-Computer Interaction*, 6, 267:1-267:28. <https://doi.org/10.1145/3555157>
- Steuperaert, D., Poels, G., & Devos, J. (2025). A Reference Model for Information Quality in an it Governance Context, *Information Systems Management*, 42(3), 346-364. <https://doi.org/10.1080/10580530.2025.2477455>
- Sveinbjarnarson, B. F., Schmitz, L., Arnardottir, E. S., & Islind, A. S. (2023). The sleep revolution platform: A dynamic data source pipeline and digital platform architecture for complex sleep data. *Current Sleep Medicine Reports*, 9(2), 91-100. <https://doi.org/10.1007/s40675-023-00252-x>
- Sveinbjarnarson, B. F., Arnardottir, E. S., & Islind, A. S. (2024). Layer Upon Layer: Developing Layered Modular Architectures for Data-Driven Health Platform. In *Digital (Eco) Systems and Societal Challenges: New Scenarios for Organizing* (pp. 91-108). Cham: Springer Nature Switzerland.
- Tracy, S. J. (2024). *Qualitative research methods: Collecting evidence, crafting analysis, communicating impact*. John Wiley & Sons.
- Wang, R. Y., & Strong, D. M. (1996). Beyond accuracy: What data quality means to data consumers. *Journal of Management Information Systems*, 12(4), 5-33.

## Appendix A

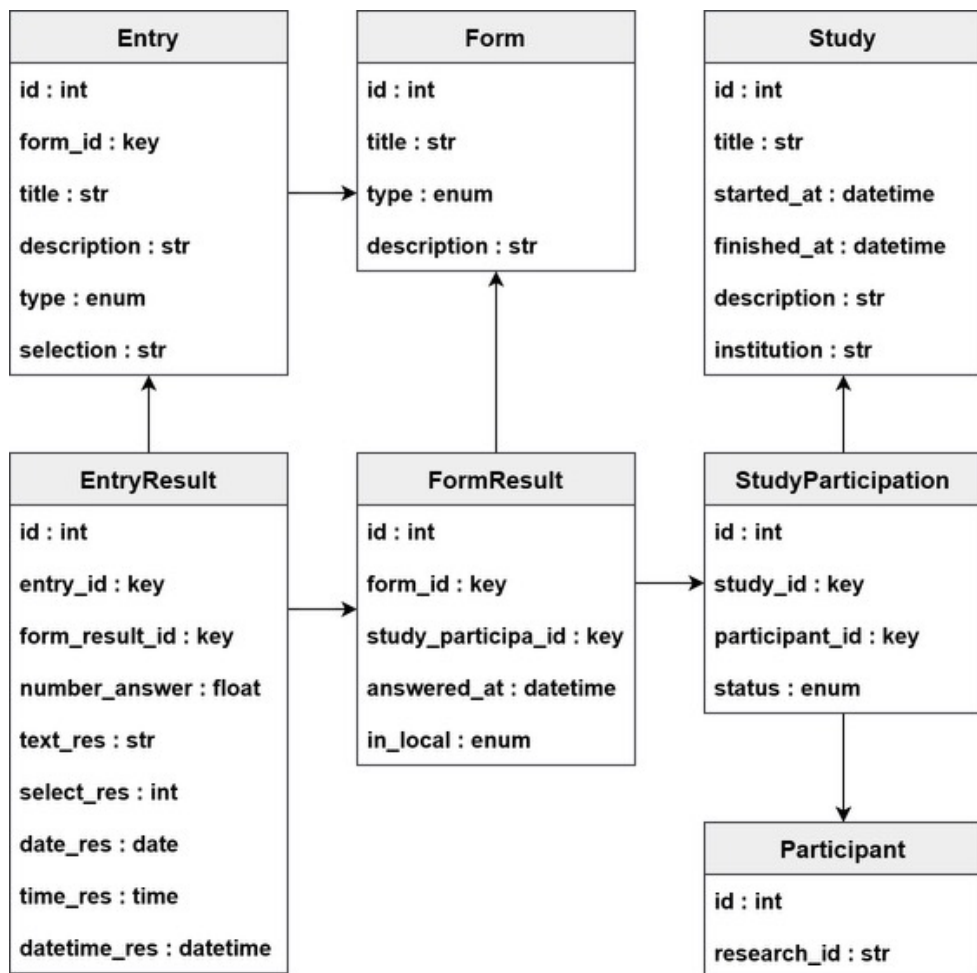


Figure A.1. Lean database structure managing which, by design, does not require additional columns or tables.

# Upload CSV File

Choose file No file chosen

Upload

Input successful!

Finished inserting, 1 rows updated, 1 new rows inserted and 0 rows you input were identical to already existing values

Starting validation for file...

Your header or column names matches study attempting input

Quick validation failed:

Your input contains cells that do not fulfill their format validations.

Excel cells this happened at are:

B2 : C2 : B3 : C3 :

the first example is: 5/1/2023

the column 'started\_at' must follow date pattern on the format YYYY-MM-DD for example '2023-01-25'

**Figure A.2.** The interface used in the test for data upload displays two example output scenarios. For space efficiency, both outputs are combined in a single image. The top blue box shows feedback when the uploaded file contains no violations, indicating the extent of database impact. The bottom blue box illustrates feedback when a violation is detected, providing an explanation of the issue to the user.

| Interview Questions |                                                                                                                                                                                                                                        |
|---------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Identifier          | Question                                                                                                                                                                                                                               |
| Q1                  | In your courses or throughout your work, have you been taught about data quality, usability, or in general, how to make data better?                                                                                                   |
| Q2                  | In your previous work, have there been data principles or some standards set before? Can you expand on that?                                                                                                                           |
| Q3                  | Do you believe that the validation system you were using now should be used for all result data in the Sleep Revolution?                                                                                                               |
| Q4                  | Did you think the validation system you were using was too complicated to use? If so, what part of it is complicated?                                                                                                                  |
| Q5                  | Do you believe the documentation was helpful? If so, what parts did the documentation help you with? Which parts were not helpful?                                                                                                     |
| Q6                  | Did you find the example input CSV helpful?                                                                                                                                                                                            |
| Q7                  | How would you describe your experience of processing the data? What aspects of it did you find nice/annoying/frustrating etc?                                                                                                          |
| Q8                  | How important is it to you that the processes around data are simple and straightforward?                                                                                                                                              |
| Q9                  | If you were in control of ensuring data quality, would you prefer a different system for validations?                                                                                                                                  |
| Q10                 | Were there any validations in the system that you encountered that you think might be unnecessary?                                                                                                                                     |
| Q11                 | Of the files you uploaded, are some files or parts of them still unclear to you?                                                                                                                                                       |
| Q12                 | How do you believe the result data and the raw data should be structured? That is, after everything has been collected, processed, and validated, how should the final data be presented so it can be used by others, in your opinion? |
| Q13                 | What do you believe are important things after data collection has finished?                                                                                                                                                           |
| Q14                 | Do you believe you could continue uploading data sources after this session?                                                                                                                                                           |
| Q15                 | Data is troublesome for projects in general when it comes to understanding it, using it, it often requires a lot of processing, contains mistakes, is not complete, and more. Why do you believe that is?                              |
| Q16                 | After a research grant period has finished, do you believe outside researchers who were not part of any of the data collections in that grant should be granted access to the data that was collected? (Can you elaborate on why?)     |
| Q17                 | Anything to add?                                                                                                                                                                                                                       |

Table A.1. The 17 interview questions we asked participants.

| Opinions: Strongly Disagree (1) to Strongly Agree (5) |                                                                                                                                                                                                                                                                                                               |
|-------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Identifier                                            | Question                                                                                                                                                                                                                                                                                                      |
| QA1                                                   | I believe manual validation (such as agreeing on standards and following them) is better in research than automatic validation (such as software or websites validating the data).                                                                                                                            |
| QA2                                                   | I believe instead of using automatic validation (such as software or websites validating the data), data users should get the data sources as they were collected, that is, data without any processing.                                                                                                      |
| QA3                                                   | I believe I can trust data that has gone through automatic validation (such as software or websites validating the data) is better than data that has not.                                                                                                                                                    |
| QA4                                                   | When I receive data that I did not collect, I do not trust it.                                                                                                                                                                                                                                                |
| QA5                                                   | When I receive data I did not collect, I have often trouble understanding what parameters/variable actually is. For example, if they are named with abbreviation or have general names.                                                                                                                       |
| QA6                                                   | When I receive data I did not collect, I often find descriptions and details (metadata) on the data sources are missing or difficult to find.                                                                                                                                                                 |
| QA7                                                   | I often find it difficult to join data sources. For example, because of missing identifiers or different formats.                                                                                                                                                                                             |
| Opinions: Not Important (1) to Very Important (5)     |                                                                                                                                                                                                                                                                                                               |
| Identifier                                            | Question                                                                                                                                                                                                                                                                                                      |
| QB1                                                   | How important is it to you that data has an existing research identifier?                                                                                                                                                                                                                                     |
| QB2                                                   | How important is it to you that data does not have duplicates when you receive it for analysis?                                                                                                                                                                                                               |
| QB3                                                   | How important is it to you that data has consistent format? For example if dates are not written as DD/MM/YY and as YYYY-MM-DD or research identifiers are not written sometimes as 'SRI0123' and sometimes 'sri0123'?                                                                                        |
| QB4                                                   | How important is it to you that you know which dataset the data source belongs to? For example, if this data was collected in one study or another?                                                                                                                                                           |
| Time Spent: Percentage                                |                                                                                                                                                                                                                                                                                                               |
| Identifier                                            | Question                                                                                                                                                                                                                                                                                                      |
| QC1                                                   | On average how much percentage of your work goes into validating data? This includes making sure your data has correct formats.                                                                                                                                                                               |
| QC2                                                   | On average how much percentage of your work goes into converting data? By converting, we mean to convert it from one data format (XML, CSV, JSON, TXT) to another in order to, for example, being able to open the data in some environment.                                                                  |
| QC3                                                   | On average how much percentage of your work goes into cleaning data? By cleaning, we mean to finding errors, make sure formats are the same (for example, all dates written with same format), applying ownership if they are missing, etc.                                                                   |
| QC4                                                   | On average how much percentage of your work goes into joining data? By joining, we mean to join data sources together so they can be worked on. For example, taking data from a text file and inputting it into an Excel file so the results can be compared, or joining two different excel sheets together. |
| QC5                                                   | On average how much of your work goes into reading, documenting, and writing?                                                                                                                                                                                                                                 |
| QC6                                                   | On average how much percentage of your work goes into meetings?                                                                                                                                                                                                                                               |
| QC7                                                   | On average how much percentage of your work goes into other things? This is the final percentage question so please make sure they add to 100%.                                                                                                                                                               |

Table A.2. Many of the survey questions we asked.

