Multimodal Technologies and Interaction

Review

Data Governance in Multimodal Behavioral Research

Zhehan Jiang [1,2], Zhengzhou Zhu [3,*] and Shucheng Pan [4,*]

[1] Institute of Medical Education, Peking University, No. 38 Xueyuan Road, Beijing 100191, China; jiangzhehan@bjmu.edu.cn
[2] National Center for Health Professions Education Development, Peking University, No. 38 Xueyuan Road, Beijing 100191, China
[3] School of Software and Microelectronics, Peking University, No.5 Yiheyuan Road, Beijing 100871, China
[4] School of Nursing, Peking University, No. 38 Xueyuan Road, Beijing 100191, China
[*] Correspondence: zhuzz@pku.edu.cn (Z.Z.); panshucheng@bjmu.edu.cn (S.P.)

Abstract: In the digital era, multimodal behavioral research has emerged as a pivotal discipline, integrating diverse data sources to comprehensively understand human behavior. This paper defines and distinguishes data governance from mere data management within this context, highlighting its centrality in assuring data quality, ethical handling, and participant protection. Through a meticulous review of the literature and empirical experience, we identify key implementation strategies and elucidate the benefits and risks of data governance frameworks in multimodal research. A demonstrative case study illustrates the practical applications and challenges, revealing enhanced data reliability and research integrity as tangible outcomes. Our findings underscore the critical need for robust data governance, pointing to future advancements in the field, including the development of adaptive governance frameworks, innovative big data analytics solutions, and user-friendly tools. These enhancements are poised to amplify the utility of multimodal data, propelling behavioral science forward.

Keywords: data governance; multimodal; behavioral; big data; analytics; healthcare; primer; framework; demonstrative

Citation: Jiang, Z.; Zhu, Z.; Pan, S. Data Governance in Multimodal Behavioral Research. *Multimodal Technol. Interact.* **2024**, *8*, 55. https://doi.org/10.3390/mti8070055

Received: 10 May 2024
Revised: 18 June 2024
Accepted: 19 June 2024
Published: 25 June 2024

1. Introduction

The advent of the digital era has transformed the landscape of research, rendering data an invaluable asset for modern society. The exponential growth of data has influenced decision-making processes across various industries and sectors, particularly within the healthcare domain [1,2]. The increasing availability of diverse data sources, such as electronic health records, social media, and wearable devices, has opened up new avenues for research and innovation [3]. In this context, multimodal behavioral research has emerged as a promising field that leverages the power of big data analytics to gain insights into human behavior and health [4,5].

Multimodal behavioral research involves the collection, integration, and analysis of data from various sources, such as physiological sensors, video recordings, and self-reports [6]. This approach enables researchers to capture a more comprehensive and nuanced understanding of human behavior by combining information from different modalities [7]. For instance, several studies investigating the relationship between stress and cardiovascular health may collect data from wearable devices measuring heart rate variability, video recordings of facial expressions, and self-reported questionnaires assessing perceived stress levels [8,9]. By integrating these diverse data streams, researchers can uncover complex patterns and interactions that may not be apparent when analyzing each modality in isolation.

However, the increasing complexity and volume of data generated in multimodal behavioral research pose significant challenges for data management and analysis [10].

The heterogeneity of data formats, the need for data harmonization, and the ethical considerations surrounding data privacy and security necessitate the development of robust data governance practices [11]. Data governance, as a multifaceted concept, addresses the strategic, tactical, and operational aspects of managing and leveraging data to achieve research goals while ensuring data quality, security, and regulatory compliance [12].

The importance of data governance in multimodal behavioral research cannot be overstated. Effective data governance practices ensure that data is accurate, reliable, and usable for research purposes [13]. This is particularly crucial in healthcare research, where data are highly decentralized across thousands of online databases, and data quality and integrity directly impact patient outcomes and public health policies [14]. Moreover, data governance frameworks help to mitigate the risks associated with data breaches, unauthorized access, and misuse, which can have severe consequences for research participants and institutions [15].

Despite the critical role of data governance in multimodal behavioral research, there is a lack of comprehensive guidelines and best practices tailored to this specific domain [11]. Existing data governance frameworks, such as the DAMA-DMBOK (Data Management Body of Knowledge) and the DGI (Data Governance Institute) Maturity Model, provide valuable insights into the general principles and processes of data governance [12,16]. However, these frameworks may not fully address the unique challenges and requirements of multimodal behavioral research, such as the integration of heterogeneous data types, the management of large-scale datasets, and the ethical considerations surrounding the use of sensitive personal information [17]. Previous studies [10,14–16,18] have primarily focused on data management strategies without fully addressing the complexities of governance in multimodal research. Methodologies often involve integrating video, audio, sensor, and survey data, yet these studies commonly report inconsistent data quality, ethical concerns, and difficulties in cross-modal analysis. Results indicate that while there is recognition of the importance of data governance, few have outlined comprehensive frameworks tailored to the specific challenges of multimodal data [19,20]. Limitations include a lack of clear guidelines for ethical data use, standardization across modalities, and efficient data processing pipelines.

To bridge this gap, there is a pressing need for a comprehensive and tailored approach to data governance in multimodal behavioral research. This approach should take into account the specific characteristics and requirements of this domain, while leveraging the best practices and lessons learned from existing data governance frameworks [11]. By developing a robust and adaptable data governance framework for multimodal behavioral research, we can ensure the quality, security, and ethical use of data, ultimately facilitating groundbreaking discoveries and advancements in the field of behavioral science.

In this paper, we aim to provide a general primer for data governance in multimodal behavioral research. We will begin by defining data governance in the context of multimodal behavioral research and differentiating it from data management. We will then discuss the various types of data governance implementation and the steps involved in establishing a data governance framework. Additionally, we will explore the benefits, disadvantages, and risks associated with data governance in multimodal behavioral research. To illustrate the practical application of data governance principles, we will present a demonstrative case study showcasing the challenges and solutions in a real-world multimodal behavioral research project. Finally, we will conclude with a discussion on the future directions and potential impact of data governance in advancing the field of multimodal behavioral research.

The workflow can be seen in Figure 1. This figure outlines the sequential stages of implementing data governance in multimodal research, starting with strategy formulation and standard setting, progressing through data acquisition and processing, ensuring ethical compliance and quality control, and concluding with secure storage and responsible sharing, all while illustrating how each step contributes to a holistic governance approach. The workflow depicted in Figure 1 encapsulates our proposed four-step process for implementing data governance in multimodal behavioral research. Step 1, Strategy Formulation and Standard Setting, involves defining the governance structure, establishing clear policies, and setting data quality and ethical guidelines. Step 2, Data Acquisition and Processing, ensures that data from various modalities are consistently collected and transformed into a compatible format for analysis. Step 3, Ethical Compliance and Quality Control, entails verifying adherence to ethical standards and applying rigorous checks for data accuracy and completeness. Lastly, Step 4, Secure Storage and Responsible Sharing, addresses the long-term preservation and controlled dissemination of data in accordance with FAIR principles. For example, in a hypothetical study on student behavior, Step 1 might include creating guidelines for collecting video, audio, and wearable sensor data. Step 2 would then involve preprocessing this data to ensure compatibility. Step 3 would confirm adherence to participant privacy rights, and Step 4 would see the archiving of this processed data for future research access, all while upholding the highest standards of data governance.

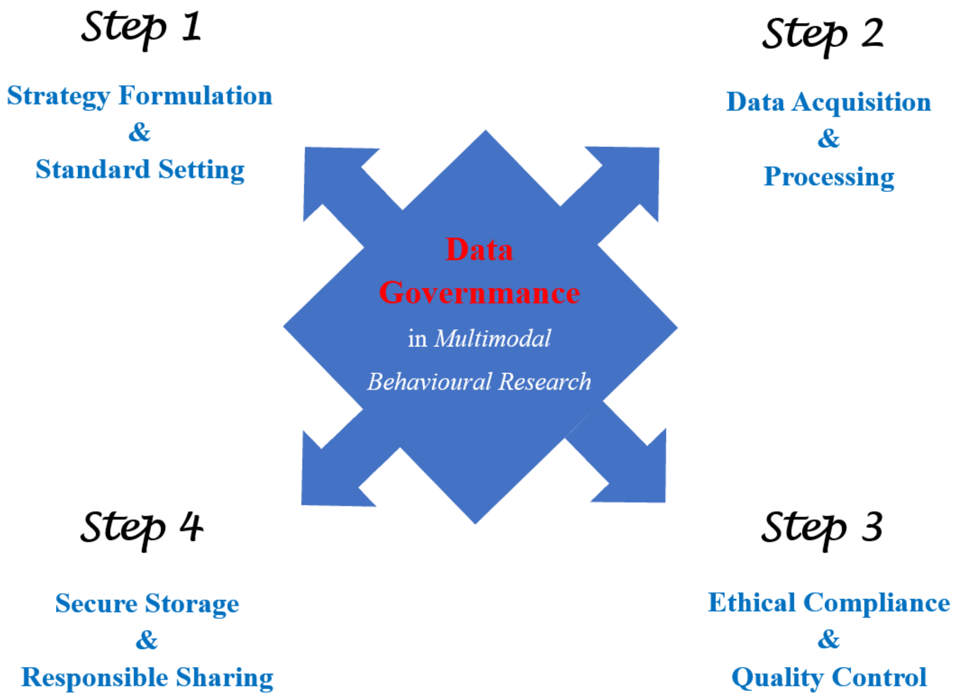

Step 1
Strategy Formulation & Standard Setting
Step 2
Data Acquisition & Processing
Data Governmance
in Multimodal Behavioural Research
Step 4
Secure Storage & Responsible Sharing
Step 3
Ethical Compliance & Quality Control

Figure 1. Workflow of Data Governance in Multimodal Behavioral Research: A Four-Step Process.

Our work fills a crucial gap by providing a comprehensive guide to data governance in multimodal behavioral research, offering a clear definition, implementation strategies, and a nuanced understanding of its benefits and risks. Our research contributes a practical framework, informed by real-world experiences, that enhances the scientific community's ability to manage and analyze multimodal data ethically and effectively. The study exemplifies the application of multimodal research methods, which are innovative in capturing a comprehensive view of complex human behaviors and experiences. This study's unique value lies in its demonstration of how proper governance fosters trust, promotes reproducibility, and accelerates scientific progress, ultimately benefiting societal well-being.

2. What Is Data Governance in Multimodal Behavioral Research?

Data governance in multimodal behavioral research ensures effective management of diverse data sources by establishing strategic oversight, clear standards, and ethical guidelines. It focuses on maintaining data quality, privacy, and security across various modalities, such as sensors and questionnaires. This framework is crucial for protecting participant rights and maintaining research integrity, distinguishing it from operational data management by emphasizing policy and accountability.

2.1. Definition of Data Governance

Data governance is a comprehensive framework that encompasses the policies, processes, and structures necessary for the effective management and utilization of an organization's data assets [13]. It involves the establishment of decision-making rights and responsibilities, the development of standards and guidelines, and the implementation of mechanisms for ensuring data quality, security, and compliance [21]. The primary goal of data governance is to maximize the value of data while minimizing the associated risks and costs [22].

The definition of data governance has evolved over time, reflecting the increasing complexity and importance of data in modern organizations. Early definitions focused primarily on the technical aspects of data management, such as data architecture and database administration [18]. However, as the volume, variety, and velocity of data have grown, the scope of data governance has expanded to include strategic, organizational, and ethical considerations [23].

From a strategic perspective, data governance is seen as a means of aligning an organization's data assets with its business objectives [21]. This involves defining the key performance indicators (KPIs) and metrics that will be used to measure the success of data initiatives, as well as establishing the governance structures and processes necessary to support data-driven decision-making [24]. For example, a healthcare organization may establish a data governance committee to oversee the development of a data warehouse that will be used to support population health management and clinical research [25].

From an organizational perspective, data governance is concerned with the roles, responsibilities, and accountability of individuals and teams involved in data management [13]. This includes defining the data stewardship roles and responsibilities, establishing clear lines of communication and collaboration between different departments and functions, and implementing mechanisms for resolving data-related disputes and conflicts [26]. For instance, a financial institution may appoint a chief data officer (CDO) to lead the development and implementation of a data governance program, working closely with business units, IT, and compliance teams [27].

From an ethical perspective, data governance is essential for ensuring that data is collected, used, and shared in a responsible and transparent manner [28]. This involves developing policies and guidelines for data privacy, security, and consent, as well as establishing the mechanisms for monitoring and auditing compliance with these policies [29]. For example, a social media company may implement a data governance framework that includes strict access controls, data anonymization techniques, and regular privacy impact assessments to protect user data from unauthorized access or misuse [30].

2.2. Differentiating Data Governance from Data Management

While data governance and data management are closely related, they are distinct concepts with different focuses and objectives. Data management refers to the technical and operational processes involved in the acquisition, storage, processing, and dissemination of data [12]. It encompasses activities such as data modeling, database design, data integration, and data quality assurance [31]. The primary goal of data management is to ensure that data is accurate, complete, consistent, and available to support business operations and decision-making [32].

In contrast, data governance is a higher-level framework that provides the strategic direction and oversight necessary for effective data management [13]. It establishes the policies, standards, and accountability mechanisms that guide data management practices and ensure that they align with an organization's overall goals and objectives [21]. While data management focuses on the day-to-day handling of data, data governance provides the overarching structure and guidance necessary for consistent and compliant data management across the organization [23].

To illustrate the difference between data governance and data management, consider the example of a healthcare organization implementing an electronic health record (EHR) system [33]. The data management team would be responsible for designing the database schema, developing data integration processes, and ensuring data quality through validation and cleansing routines. The data governance team, on the other hand, would be responsible for defining the data ownership and access policies, establishing data standards and definitions, and ensuring compliance with privacy and security regulations such as HIPAA (Health Insurance Portability and Accountability Act) [13].

Another key difference between data governance and data management is the scope and level of involvement of different stakeholders. Data management is primarily the responsibility of IT and data management professionals, who have the technical expertise necessary to design, implement, and maintain data systems [31]. Data governance, on the other hand, requires the involvement and collaboration of a wide range of stakeholders, including business leaders, legal and compliance experts, and data users [21]. This is because data governance decisions often have significant implications for an organization's strategy, operations, and risk management [22].

2.3. Data Governance in Multimodal Behavioral Research

Multimodal behavioral research involves the collection, integration, and analysis of data from multiple sources and modalities, such as physiological sensors, video recordings, and self-report questionnaires [6]. This approach offers several advantages over traditional single-modality research, including increased ecological validity, an improved understanding of complex behaviors, and the ability to capture both objective and subjective aspects of the human experience [34]. However, the complexity and heterogeneity of multimodal data also pose significant challenges for data governance [11].

One of the key challenges in governing multimodal behavioral data is ensuring data quality and consistency across different modalities and sources [34]. Each modality may have its own unique data formats, sampling rates, and measurement scales, making it difficult to integrate and compare data from different sources [6]. Moreover, the quality and reliability of data may vary depending on the specific devices, sensors, or instruments used, as well as the data collection protocols and procedures followed [11]. To address these challenges, data governance frameworks for multimodal behavioral research need to establish clear data standards and definitions, as well as processes for data harmonization and integration.

Another important consideration in governing multimodal behavioral data is ensuring data privacy and security [17]. Multimodal data often includes sensitive personal information, such as biometric data, location data, and audio/video recordings, which may be subject to strict privacy regulations and ethical guidelines [15]. Data governance frameworks need to establish clear policies and procedures for obtaining informed consent, protecting participant confidentiality, and securing data from unauthorized access or breaches [11]. This may involve implementing access controls, encryption mechanisms, and data anonymization techniques, as well as regular audits and risk assessments [17].

In addition to data quality and privacy, data governance in multimodal behavioral research also needs to consider the ethical implications of data collection, use, and sharing [35]. Multimodal data may reveal sensitive information about individuals' behaviors, preferences, and mental states, raising concerns about potential misuse or exploitation [17]. Data governance frameworks need to establish clear guidelines for the responsible and transparent use of multimodal data, taking into account the potential risks and benefits to participants, researchers, and society as a whole [35]. This may involve developing data sharing agreements, establishing the oversight committees, and engaging in ongoing dialogue with research participants and other stakeholders [11].

To illustrate the importance of data governance in multimodal behavioral research, consider the example of a study investigating the relationship between stress, physical activity, and sleep quality. The research team may collect data from wearable devices (e.g., accelerometers and heart rate monitors), smartphone apps (e.g., sleep and mood trackers), and daily questionnaires [34]. To ensure data quality and consistency, the data governance framework would need to establish data standards and protocols for each modality, such as minimum sampling rates, calibration procedures, and data formatting requirements [6]. The framework would also need to include processes for data cleaning, validation, and integration, such as algorithms for detecting and correcting sensor artifacts or missing data [11].

To protect participant privacy and confidentiality, the data governance framework would need to establish clear policies and procedures for informed consent, data access, and data sharing [17]. This may involve obtaining separate consent for each modality, restricting access to sensitive data to authorized personnel only, and implementing secure data storage and transmission protocols [15]. The framework would also need to include mechanisms for participant feedback and control, such as the ability to withdraw consent or request data deletion [35].

Finally, to address the ethical implications of multimodal data collection and use, the data governance framework would need to establish guidelines for the responsible conduct of research, taking into account the potential risks and benefits to participants and society [11]. This may involve developing data sharing agreements that specify the conditions under which data can be accessed and used by third parties, establishing independent oversight committees to review research proposals and monitor ongoing studies, and engaging in regular communication and outreach with research participants and other stakeholders [35].

That said, data governance is a critical component of multimodal behavioral research, ensuring that data is collected, managed, and used in a consistent, secure, and ethical manner. By establishing clear policies, standards, and accountability mechanisms, data governance frameworks can help researchers navigate the complex challenges of multimodal data integration and analysis, while also protecting the rights and interests of research participants. As the field of multimodal behavioral research continues to evolve, it will be essential to develop and refine data governance practices such that they can keep pace with the rapidly changing technological and regulatory landscape.

3. How Can Data Governance Be Implemented in Multimodal Behavioral Research?

The implementation of data governance in multimodal behavioral research involves a series of methodical and strategic actions that cater to the unique requirements of multimodal data. Effective implementation not only streamlines data management practices but also aligns with the overarching goals of the research organization. We will discuss them by the following: types of implementations, typical implementation steps, as well as software and toolkits.

3.1. Types of Implementations

Research organizations can implement data governance in various ways, adopting centralized, decentralized, or hybrid models depending on their organizational structure, culture, and maturity [36]. Centralized models feature a single authoritative body that governs data across the organization, while decentralized models delegate governance responsibilities to individual research teams or departments [37]. Hybrid models combine the strengths of both, ensuring that universal standards are maintained while allowing flexibility at the local level [38].

Centralized data governance models feature a single authoritative body that is responsible for governing data across the entire organization [37]. This central authority, often in the form of a data governance council or committee, sets the policies, standards, and procedures that guide data management practices, and ensures compliance through regular monitoring and auditing [21]. The centralized approach offers several advantages, including consistency and standardization by establishing a single set of policies and standards that ensure data is managed consistently across the organization; efficiency and economies of scale by allowing for the pooling of resources and expertise, reducing duplication of effort and enabling the sharing of best practices [13]; and stronger oversight and accountability with a single point of control, making it easier to monitor compliance, enforce policies, and hold individuals and teams accountable for their data management practices [22]. However, centralized governance also has some drawbacks, such as reduced flexibility and agility, as centralized policies and standards may not always fit the specific needs or requirements of individual research teams or projects, leading to delays or workarounds [37]; limited local ownership and engagement, as centralized governance may be perceived as a top-down imposition, leading to resistance or a lack of buy-in from researchers and data users [23]; and potential bottlenecks and bureaucracy, as a single decision-making body may create bottlenecks or bureaucratic hurdles that slow down data access and utilization [38].

Decentralized data governance models delegate governance responsibilities to individual research teams or departments, allowing them to develop their own policies, standards, and procedures based on their specific needs and requirements [37]. This approach offers several benefits, including flexibility and adaptability, enabling research teams to tailor their data management practices to their specific research questions, methodologies, and data types [13]; local ownership and engagement, fostering a sense of ownership and engagement among researchers and data users, leading to higher levels of compliance and buy-in [23]; and faster decision-making and innovation, allowing research teams to make decisions and implement changes more quickly, without having to navigate complex bureaucratic processes [38]. However, decentralized governance also has some limitations, such as lack of consistency and standardization, potentially leading to inconsistencies or conflicts in data management practices across the organization [21]; duplication of effort and inefficiencies, resulting in teams or departments reinventing the wheel or duplicating efforts, leading to inefficiencies and wasted resources [13]; and reduced oversight and accountability, potentially leading to gaps or weaknesses in data security, privacy, or quality without a central authority to monitor compliance and enforce policies [22].

Hybrid data governance models combine elements of both centralized and decentralized approaches, seeking to balance the benefits of consistency and standardization with the need for flexibility and local ownership [38]. In a hybrid model, a central authority sets the overall policies, standards, and guidelines for data governance, while individual research teams or departments are responsible for implementing and adapting these policies to their specific contexts [37]. The hybrid approach offers several advantages, such as balancing global and local needs by ensuring that universal standards and best practices are maintained across the organization, while also allowing for local variations and adaptations based on specific research requirements [23]; fostering collaboration and knowledge sharing by creating a shared framework for data governance that encourages collaboration and knowledge sharing between different teams and departments, leading to improved efficiency and innovation [13]; and enabling continuous improvement and learning by allowing for the identification and dissemination of best practices and lessons learned across the organization [38]. However, hybrid governance also has some challenges, such as complexity and coordination, requiring careful coordination and communication between the central authority and local teams, which can be complex and time-consuming [21]; potential for conflict and confusion, as multiple levels of decision-making and implementation may lead to conflicts or confusion about roles, responsibilities, and accountability [37]; and resource and capability requirements, necessitating significant investments in resources, skills, and capabilities, both at the central and local levels, to ensure effective implementation and ongoing management [23].

A large research university with multiple departments and research centers can illustrate the differences between centralized, decentralized, and hybrid data governance models. In a centralized model, the university's central IT department would develop and enforce data governance policies and standards across all research units, with limited local input or flexibility. In a decentralized model, each department or research center would develop its own data governance practices, with little or no central coordination or oversight. In a hybrid model, the central IT department would establish core policies, standards, and guidelines for data governance, covering key areas such as data security, privacy, quality, and metadata management, in consultation with representatives from different research units. Each department or research center would then implement these policies in their specific contexts, adapting them to fit their research workflows, data types, and local requirements. The central IT department would provide guidance, support, and oversight to ensure that the local implementations align with the overall data governance framework while facilitating knowledge sharing and collaboration across research units.

For example, the central IT department may require all research data to be stored in a secure, centralized repository with access controls and backup procedures. The psychology department may implement this policy by setting up a departmental data repository with additional access controls and data sharing agreements based on their sensitive research data. The biology department may use a cloud-based data storage platform that integrates with their existing research workflows and data analysis tools. The central IT department would ensure both implementations meet core security and privacy requirements while providing guidance on best practices for data management and sharing.

The choice of data governance implementation model depends on factors such as the organization's size and complexity, the nature and sensitivity of the research data, the level of maturity and capability of different research units, and the balance between consistency and flexibility. Centralized models offer strong oversight and standardization but may limit local ownership and agility. Decentralized models enable flexibility and innovation but may lead to inconsistencies and inefficiencies. Hybrid models balance these trade-offs by combining central guidance with local adaptation but require careful coordination and resource investments. The most effective data governance model aligns with the organization's strategic goals, culture, and values, enabling researchers to leverage data assets for maximum impact while minimizing risks and costs.

3.2. Typical Implementation Steps

In multimodal behavioral research, effective data governance is achieved by forming a Data Governance Committee (DGC) to oversee the development of policies and procedures, while also defining roles and ensuring accountability. Regular assessments and audits help identify gaps and refine practices. Education and communication are key to building a data-literate culture, while continuous evaluation ensures that the governance framework remains effective and relevant. Typical implementation steps contain seven steps as follows: (1) Establish a Data Governance Committee (DGC); (2) assess the current state; (3) develop policies and procedures; (4) define roles and responsibilities; (5) implement metrics and monitoring; (6) foster education and communication; and (7) continuously evaluate and improve. The detailed actions for each step will be illustrated later in the demonstrative case in multimodal behavioral research.

3.3. Software and Toolkits

Since data governance in multimodal behavioral research involves managing, integrating, and ensuring the quality and security of data from various sources, there is no one-size-fits-all tool for this domain; however, several software and toolkits can be employed to support different aspects of data governance.

- Data Management Platforms (DMPs): Tools like Cloudera, Snowflake, or Microsoft Azure Synapse Analytics can be used to store, manage, and process large volumes of multimodal data.
- Data Integration Tools: For integrating data from different modalities, you might use ETL (Extract, Transform, Load) tools like Talend, Informatica PowerCenter, or Apache NiFi which help in cleaning, formatting, and combining datasets.
- Metadata Management: Alation, Collibra, or IBM Watson Knowledge Catalog provide solutions for managing metadata, defining data lineage, and maintaining a data glossary which is crucial in understanding the context of multimodal data.
- Data Quality Assurance: Tools like Trillium Software (Trillium Software, Inc., Burlington, MA, USA.), Talend Data Quality (Talend. Inc., San Mateo, CA, USA), or Open-

Refine (ver 3.8.1) can ensure data accuracy, completeness, and consistency across multiple datasets.

- Research Workflow and Project Management: Platforms like REDCap (Research Electronic Data Capture) (REDCap Inc., Fort Lauderdale, FL, USA), Qualtrics (Qualtrics, Inc., Provo, UT, USA.), or Labguru (BioData Inc., Westborough, MA, USA) can assist with study design, consent management, and data collection.
- Privacy and Security: Solutions like AWS Identity and Access Management (IAM), Azure Active Directory, or HashiCorp Vault can be used to enforce access control and protect sensitive research data.
- Behavioral Analysis Tools: Specific to behavioral research, tools like ELAN (ELAN Corporation, Matsumoto, Japan) (for annotating videos), R or Python libraries (e.g., OpenFace for facial expression analysis, MNE-Python for EEG/MEG data processing), and MATLAB toolboxes (e.g., Psychtoolbox for stimulus presentation and response logging).
- Compliance and Consent Management: Platforms like Consentric or Ethical Intelligence can help manage informed consent and ensure compliance with regulations like GDPR and HIPAA.
- Data Sharing and Archiving: Dataverse, OSF (Open Science Framework) (Center for Open Science, Charlottesville, VA, USA.), or Zenodo can be used for sharing and archiving research data according to FAIR principles (Findable, Accessible, Interoperable, and Reusable).

4. Complexities to Consider When Governing Data in Multimodal Behavioral Research

4.1. Benefits of Data Governance in Multimodal Behavioral Research

Effective data governance in multimodal behavioral research can result in numerous gains, including improved data quality, reduced redundancy, increased transparency, and better risk management [39]. Enhanced data quality leads to more reliable analytics and insights, which can drive research discoveries and improve the reproducibility of findings [40]. Furthermore, sound data governance practices can bolster trust in data-driven research products and services, thereby fostering collaboration and funding opportunities [41].

4.2. Disadvantages and Drawbacks

Despite the benefits, implementing data governance in multimodal behavioral research comes with challenges and drawbacks [10]. It can be resource-intensive, requiring significant investment in personnel, technology, and time. Additionally, data governance initiatives may face resistance due to cultural barriers, conflicting priorities, and the complexity of integrating diverse multimodal datasets. There may also be a perceived slowdown in data access and agility, especially if the governance processes are overly restrictive or bureaucratic [42].

4.3. Risks to Consider

Data governance in multimodal behavioral research must address a range of risks, such as regulatory non-compliance penalties, reputational damage from data breaches or misuse, and legal exposure due to inaccurate or incomplete data [43]. Furthermore, the rapidly evolving digital landscape introduces new risks, including the misalignment of governance policies with the emerging technologies used in multimodal data collection and analysis [44].

5. A Demonstrative Case in Multimodal Behavioral Research

A multidisciplinary team of psychologists, neuroscientists, and educational researchers collaborate on a large-scale longitudinal study to investigate the complex interplays between sleep patterns, stress hormone levels, academic workload, and student mental health. The study targets a diverse group of thousands of university students across different campuses, divided equally between the freshman and senior cohorts. The researchers utilize a combination of multimodal data sources to capture a holistic picture of the student experience.

5.1. Sources of Data

In the study, the students don wrist-worn actigraphy devices to monitor sleep duration, quality, and circadian rhythms. They provide daily saliva samples to measure cortisol levels, which serve as an indicator of stress response. Weekly online self-reports are completed by the students to track their sleep habits, stressors, mood, and study load. Academic performance data, including grades and class attendance rates, are collected from institutional records. Furthermore, the students' Twitter feeds and Instagram posts are analyzed to gauge their social interactions and mood fluctuations.

5.2. Problem Identification

The study faces several challenges in data collection and research design. The actigraphy devices from different manufacturers can exhibit variations in sensitivity and noise thresholds, leading to inconsistent sleep data quality. Similarly, salivary cortisol measurements are affected by laboratory differences and timing inconsistencies in sample collection. The self-report measures used across the various campuses are not uniform, with differences in survey wording and response scales that hinder direct comparisons of data. Privacy and consent issues arise from the collection and aggregation of social media data, requiring robust consent protocols and anonymization procedures to address privacy concerns. Additionally, the observational nature of the study, lacking randomization in device assignment and sample collection, may introduce confounding variables that could affect the reliability of the findings.

5.3. Necessity for Data Governance

Without a comprehensive data governance strategy, inconsistencies in data collection, processing, and analysis can lead to flawed interpretations, invalid statistical comparisons, and compromised study conclusions. To ensure valid findings, researchers must develop a coherent data governance plan that includes standardizing actigraphy and cortisol assay protocols across all participating sites to ensure uniformity. Additionally, self-report measures should be harmonized by revising and aligning survey instruments to a central metric system, which will facilitate direct comparisons across different datasets. The plan should also establish rigorous ethical guidelines for accessing and processing social media data, which includes implementing transparent consent processes and ensuring secure data storage. Finally, integrating data from disparate sources into a single, structured database that adheres to the FAIR principle (Findable, Accessible, Interoperable, and Reusable) will further enhance the integrity and usability of the research data.

5.4. Methods

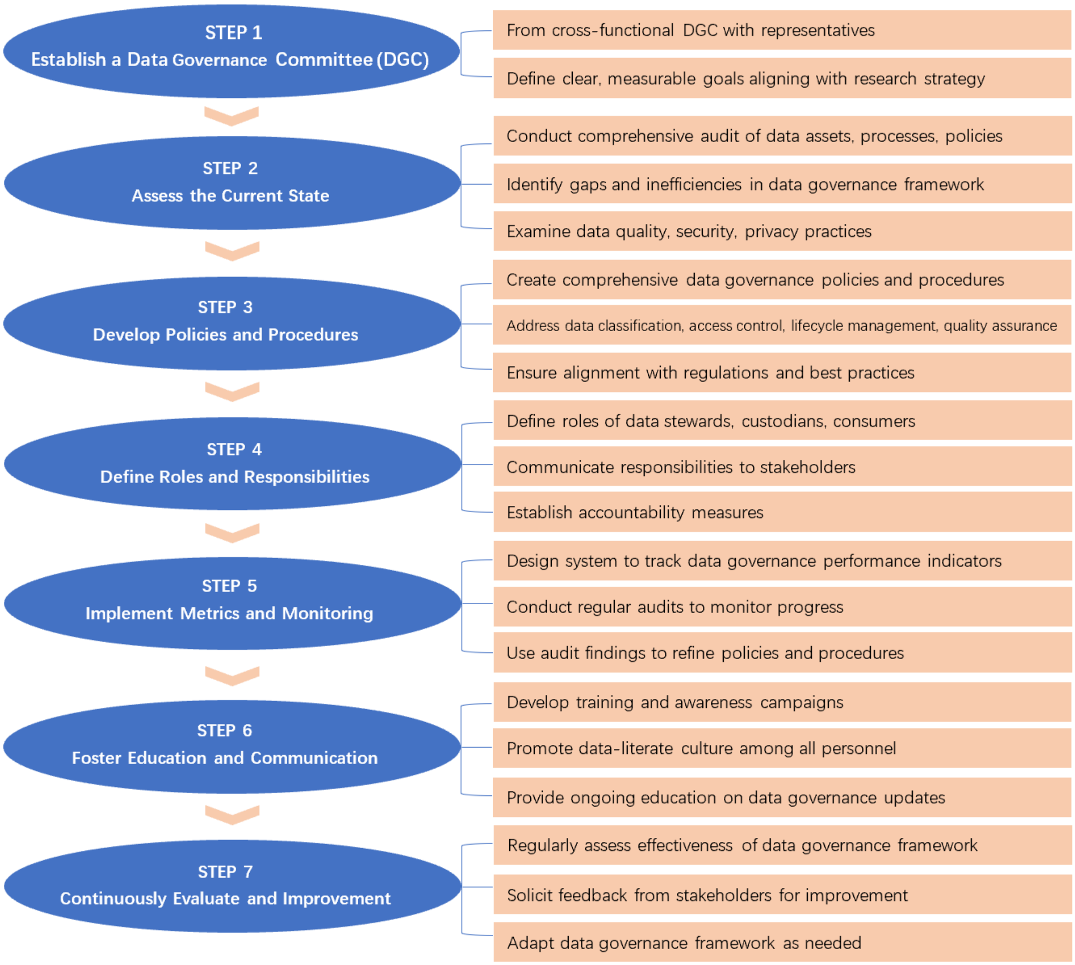
Researchers need to implement a unified data architecture that integrates the multimodal data streams into a single analytic framework, applying common metadata tags and data dictionaries to ensure consistency across different types of data. They establish data validation routines and automated quality checks during the data entry and preprocessing stages to maintain high standards of data integrity. Additionally, statistical harmonization techniques such as propensity score matching and instrumental variable analysis are applied to address potential selection biases and confounders in observational data, enhancing the reliability of the research findings.

5.5. Analysis and Evaluation

Embarking on a multidimensional exploration of the complex dynamics between sleep, stress, and academic achievement, our study is meticulously designed with a Mixed-Methods Approach, Longitudinal Analysis, and Data Governance Impact Assessment to ensure a robust and reliable examination of the subject matter.

- Mixed-Methods Approach: The study uses both quantitative and qualitative methods to triangulate findings from the different data modalities.
- Longitudinal Analysis: Researchers perform repeated measures ANOVA and multi-level modeling to examine changes in sleep patterns, stress, and academic performance over time.
- Data Governance Impact Assessment: Throughout the study, the effectiveness of the data governance strategy is continually evaluated by monitoring data completeness, internal consistency, and the ability to detect expected relationships among variables.

The implementation of a robust data governance strategy in this multimodal research design proved essential to overcoming the challenges in data collection, harmonization, and analysis. The detailed implementations are listed in Figure 2. Through a meticulous approach to standardizing the protocols, managing ethical considerations, and ensuring data quality, the study was able to generate meaningful insights into the complex interplays between sleep, stress, and academic performance among the university students. This demonstration highlights the critical role of data governance in enhancing the validity, reproducibility, and translational value of research findings in the realm of multimodal behavioral studies.


STEP 1
Establish a Data Governance Committee (DGC)
- From cross-functional DGC with representatives
- Define clear, measurable goals aligning with research strategy

STEP 2
Assess the Current State
- Conduct comprehensive audit of data assets, processes, policies
- Identify gaps and inefficiencies in data governance framework
- Examine data quality, security, privacy practices

STEP 3
Develop Policies and Procedures
- Create comprehensive data governance policies and procedures
- Address data classification, access control, lifecycle management, quality assurance
- Ensure alignment with regulations and best practices

STEP 4
Define Roles and Responsibilities
- Define roles of data stewards, custodians, consumers
- Communicate responsibilities to stakeholders
- Establish accountability measures

STEP 5
Implement Metrics and Monitoring
- Design system to track data governance performance indicators
- Conduct regular audits to monitor progress
- Use audit findings to refine policies and procedures

STEP 6
Foster Education and Communication
- Develop training and awareness campaigns
- Promote data-literate culture among all personnel
- Provide ongoing education on data governance updates

STEP 7
Continuously Evaluate and Improvement
- Regularly assess effectiveness of data governance framework
- Solicit feedback from stakeholders for improvement
- Adapt data governance framework as needed

Figure 2. Seven-step Implementation for the Demonstrative Case.

6. Discussion

The DAMA-DMBOK Functional Framework [12] delineates ten primary data management functions that contribute to a robust data governance environment, including Data Governance, Data Architecture Management, Data Development, Database Operations Management, Data Security Management, Reference and Master Data Management, Data

Warehousing and Business Intelligence Management, Document and Content Management, Metadata Management, and Data Quality Management. Each function plays a critical role in mitigating the risks associated with data governance while optimizing the benefits in multimodal behavioral research.

When embarking on a data governance journey in multimodal behavioral research, organizations should adapt these frameworks to their unique contexts, taking into account the tradeoffs inherent in each decision. For example, investing in a comprehensive data quality program could initially increase costs, but ultimately lead to cost savings and enhanced research productivity by reducing errors and duplicative efforts [45]. Moreover, a study on the implementation of data governance in a multimodal behavioral research setting [46] revealed that despite the initial hurdles, the successful integration of data governance principles resulted in a significant reduction in data-related errors and improved research outcomes.

Data governance in multimodal behavioral research is not a one-size-fits-all endeavor; rather, it requires careful planning, customization, and continuous improvement to strike a balance between control and agility, between risk mitigation and innovation. By engaging in thoughtful and strategic data governance, research organizations can unlock the true potential of their multimodal data assets and pave the way for groundbreaking discoveries and advancements in the field of behavioral science.

7. Conclusions

In conclusion, the multifaceted nature of multimodal behavioral research necessitates a sophisticated approach to data governance that is both robust and adaptable. As the volume and variety of data continues to expand, the need for effective data governance practices becomes increasingly paramount. Future work in this domain should focus on refining data governance frameworks to accommodate the evolving technological landscape, enhancing the interoperability of diverse data sources, and ensuring compliance with emerging regulations and ethical standards. Additionally, research should explore innovative solutions for managing the complexities of big data analytics, including advanced data integration techniques, artificial intelligence for data quality assurance, and machine learning algorithms for predictive analytics. Furthermore, the development of user-friendly tools and platforms that facilitate data governance tasks for researchers across different disciplines will be crucial. By prioritizing these areas of future work, the research community can harness the full potential of multimodal data to drive impactful insights and contribute to the advancement of behavioral science.

Author Contributions: Conceptualization, Z.J. and S.P.; investigation, Z.J.; resources, Z.J.; writing—original draft preparation, Z.J.; writing—review and editing, Z.J., Z.Z. and S.P.; visualization, S.P.; project administration, Z.J.; funding acquisition, Z.J. and Z.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Humanities and Social Sciences Research Planning Fund Project of The Ministry Of Education, grant number 23YJA880091; National Natural Science Foundation of China for Young Scholars, grant number 72104006; Key Laboratory of Digital Educational Publishing Technology and Standards and Digital Education Research Institute of the Peoples Education Press, grant number RJB0623001; and Peking University Health Science Center Medical Education Research Funding Project, grant number 2023YB24 and 2022YB41.

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

1. Lysaght, T.; Lim, H.Y.; Xafis, V.; Ngiam, K.Y. AI-Assisted Decision-making in Healthcare: The Application of an Ethics Framework for Big Data in Health and Research. *Asian Bioeth. Rev.* **2019**, *11*, 299–314. [CrossRef] [PubMed]
2. Asthana, S.; Mukherjee, S.; Phelan, A.L.; Standley, C.J. Governance and Public Health Decision-Making during the COVID-19 Pandemic: A Scoping Review. *Public Health Rev.* **2024**, *45*, 1606095. [CrossRef]

3. Brown, P.A.; Anderson, R.A. A methodology for preprocessing structured big data in the behavioral sciences. *Behav. Res. Methods* **2023**, *55*, 1818–1838. [CrossRef] [PubMed]
4. Elshawi, R.; Sakr, S.; Talia, D.; Trunfio, P. Big Data Systems Meet Machine Learning Challenges: Towards Big Data Science as a Service. *Big Data Res.* **2018**, *14*, 1–11. [CrossRef]
5. Alvarez-Romero, C.; Martínez-García, A.; Bernabeu-Wittel, M.; Parra-Calderón, C.L. Health data hubs: An analysis of existing data governance features for research. *Health Res. Policy Syst.* **2023**, *21*, 70. [CrossRef] [PubMed]
6. Lahat, D.; Adali, T.; Jutten, C. Multimodal Data Fusion: An Overview of Methods, Challenges, and Prospects. *Proc. IEEE* **2015**, *103*, 1449–1477. [CrossRef]
7. Baltrušaitis, T.; Ahuja, C.; Morency, L.P. Multimodal Machine Learning: A Survey and Taxonomy. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *41*, 423–443. [CrossRef]
8. Bayoumy, K.; Gaber, M.; Elshafeey, A.; Mhaimeed, O.; Dineen, E.H.; Marvel, F.A.; Martin, S.S.; Muse, E.D.; Turakhia, M.P.; Tarakji, K.G.; et al. Smart wearable devices in cardiovascular care: Where we are and how to move forward. *Nat. Rev. Cardiol.* **2021**, *18*, 581–599. [CrossRef]
9. Prieto-Avalos, G.; Cruz-Ramos, N.A.; Alor-Hernández, G.; Sánchez-Cervantes, J.L.; Rodríguez-Mazahua, L.; Guarneros-Nolasco, L.R. Wearable Devices for Physical Monitoring of Heart: A Review. *Biosensors* **2022**, *12*, 292. [CrossRef]
10. Mangaroska, K.; Martinez-Maldonado, R.; Vesin, B.; Gašević, D. Challenges and opportunities of multimodal data in human learning: The computer science students' perspective. *J. Comput. Assist. Learn.* **2021**, *37*, 1030–1047. [CrossRef]
11. Choudhury, S.; Fishman, J.R.; McGowan, M.L.; Juengst, E.T. Big data, open science and the brain: Lessons learned from genomics. *Front. Hum. Neurosci.* **2014**, *8*, 239. [CrossRef] [PubMed]
12. DAMA International. *DAMA-DMBOK Revised Edition*, 2nd ed.; Technics Publications: Denville, NJ, USA, 2024; pp. 100–121.
13. Khatri, V.; Brown, C.V. Designing data governance. *Commun. ACM* **2010**, *53*, 148–152. [CrossRef]
14. McMurry, J.A.; Juty, N.; Blomberg, N.; Burdett, T.; Conlin, T.; Conte, N.; Courtot, M.; Deck, J.; Dumontier, M.; Fellows, D.K.; et al. Identifiers for the 21st century: How to design, provision, and reuse persistent identifiers to maximize utility and impact of life science data. *PLoS Biol.* **2017**, *15*, e2001414. [CrossRef]
15. Ferretti, A.; Ienca, M.; Sheehan, M.; Blasimme, A.; Dove, E.S.; Farsides, B.; Friesen, P.; Kahn, J.; Karlen, W.; Kleist, P.; et al. Ethics review of big data research: What should stay and what should be reformed? *BMC Med. Ethics.* **2021**, *22*, 51. [CrossRef] [PubMed]
16. The DGI Data Governance Framework. Available online: https://datagovernance.com/the-dgi-data-governance-framework/ (accessed on 1 January 2020).
17. Schwartz, P.H.; Caine, K.; Alpert, S.A.; Meslin, E.M.; Carroll, A.E.; Tierney, W.M. Patient Preferences in Controlling Access to Their Electronic Health Records: A Prospective Cohort Study in Primary Care. *J. Gen. Intern. Med.* **2017**, *30*, 25–30. [CrossRef]
18. Abraham, R.; Schneider, J.; Vom Brocke, J. Data governance: A conceptual framework, structured review, and research agenda. *Int. J. Inf. Manag.* **2019**, *49*, 424–438. [CrossRef]
19. Alwahaby, H.; Cukurova, M.; Papamitsiou, Z.; Giannakos, M. *The Multimodal Learning Analytics Handbook*; Springer: Cham, Switzerland, 2022; pp. 289–325.
20. Norris, S. *Systematically Working with Multimodal Data: Research Methods in Multimodal Discourse Analysis*; John Wiley & Sons: New York, NY, USA, 2019; pp. 159–195.
21. Rahimi, F.; Kaleibar, F.J.; Feizi, F.; Nia, A.H.; Kashfi, H. Navigating Data Governance in the Telecom Industry. In Proceedings of the 7th Iranian Conference on Advances in Enterprise Architecture (ICAEA), Tehran, Iran, 15–16 November 2023.
22. Alhassan, I.; Sammon, D.; Daly, M. Data governance activities: A comparison between scientific and practice-oriented literature. *J. Enterp. Inf. Manag.* **2018**, *31*, 300–316. [CrossRef]
23. Alhassan, I.; Sammon, D.; Daly, M. Data governance activities: An analysis of the literature. *J. Decis. Syst.* **2016**, *25* (Suppl. S1), 64–75. [CrossRef]
24. Marcucci, S.; Alarcón, N.G.; Verhulst, S.G.; Wüllhorst, E. Informing the Global Data Future: Benchmarking Data Governance Frameworks. *Data Policy* **2023**, *5*, e30. [CrossRef]
25. Holmes, E.A.; Craske, M.G.; Graybiel, A.M. Psychological treatments: A call for mental-health science. *Nature* **2014**, *511*, 287–289. [CrossRef]
26. Bergren, M.D. Data Governance and Stewardship. *NASN Sch Nurse.* **2019**, *34*, 149–151. [CrossRef] [PubMed]
27. Pandey, N.; Dé, R.; Ravishankar, M. Improving the governance of information technology: Insights from the history of Internet governance. *J. Inf. Technol.* **2022**, *37*, 266–287. [CrossRef]
28. Floridi, L.; Taddeo, M. What is data ethics? Philosophical Transactions of the Royal Society A: Mathematical. *Phys. Eng. Sci. Med.* **2016**, *374*, 20160360.
29. Colesky, M.; Hoepman, J.-H.; Hillen, C. A Critical Analysis of Privacy Design Strategies. In Proceedings of the 2016 IEEE Security and Privacy Workshops (SPW), San Jose, CA, USA, 22–26 May 2016; pp. 33–40.
30. Michota, A.; Katsikas, S. Towards improving existing online social networks' privacy policies. *Int. J. Inf. Priv. Secur. Integr.* **2018**, *3*, 209–229.
31. Berson, A.; Dubov, L. *Master Data Management and Data Governance*, 2nd ed.; McGraw-Hill: New York, NY, USA, 2011; pp. 153–180.
32. Ram, J.; Afridi, N.K.; Khan, K.A. Adoption of Big Data analytics in construction: Development of a conceptual model. *Built Environ. Proj. Asset Manag.* **2019**, *9*, 564–579. [CrossRef]

33. Sunyaev, A.; Dehling, T.; Taylor, P.L.; Mandl, K.D. Availability and quality of mobile health app privacy policies. *J. Am. Med. Inform. Assoc.* **2015**, *22*, e28–e33. [CrossRef] [PubMed]
34. Xie, C.; Gao, J.; Tao, C. Big Data Validation Case Study. In Proceedings of the 2017 IEEE Third International Conference on Big Data Computing Service and Applications (BigDataService), San Francisco, CA, USA, 6–9 April 2017; pp. 281–286.
35. Mittelstadt, B.D.; Floridi, L. The Ethics of Big Data: Current and Foreseeable Issues in Biomedical Contexts. *Sci. Eng. Ethic* **2015**, *22*, 303–341. [CrossRef] [PubMed]
36. Ahmed, K.; Sachindra, D.; Shahid, S.; Iqbal, Z.; Nawaz, N.; Khan, N. Multi-model ensemble predictions of precipitation and temperature using machine learning algorithms. *Atmospheric Res.* **2019**, *236*, 104806. [CrossRef]
37. Archer, J.; Stevenson, L.; Couzens, J.; Ripley, E. Connecting patient experience, leadership, and the importance of involvement, information, and empathy in the care process. *Healthc. Manag. Forum* **2018**, *31*, 252–255. [CrossRef]
38. Lee, M.D.; Criss, A.H.; Devezer, B.; Donkin, C.; Etz, A.; Leite, F.P.; Matzke, D.; Rouder, J.N.; Trueblood, J.S.; White, C.N.; et al. Robust Modeling in Cognitive Science. *Comput. Brain Behav.* **2019**, *2*, 141–153. [CrossRef]
39. Strong, D.M.; Lee, Y.W.; Wang, R.Y. Data quality in context. *Commun. ACM* **1997**, *40*, 103–110. [CrossRef]
40. Braithwaite, J.; Herkes, J.; Ludlow, K.; Testa, L.; Lamprell, G. Association between organisational and workplace cultures, and patient outcomes: Systematic review. *BMJ Open* **2018**, *7*, e017708. [CrossRef] [PubMed]
41. Harris, M.A.; Levy, A.R.; Teschke, K.E. Personal privacy and public health: Potential impacts of privacy legislation on health research in Canada. *Can. J. Public Health* **2019**, *99*, 293–296. [CrossRef] [PubMed]
42. Willcocks, L.; Lacity, M. *Service Automation: Robots and the Future of Work*; Steve Brookes Publishing: Warwickshire, UK, 2016; pp. 203–233.
43. Redman, T.C. Data's credibility problem. *Harv. Bus. Rev.* **2019**, *91*, 84–88.
44. De Sousa, W.G.; de Melo, E.R.P.; Bermejo, P.H.D.S.; Farias, R.A.S.; Gomes, A.O. How and where is artificial intelligence in the public sector going? A literature review and research agenda. *Gov. Inf. Q.* **2020**, *36*, 101392. [CrossRef]
45. Haverila, M.; Haverila, K.; Gani, M.O.; Mohiuddin, M. The relationship between the quality of big data marketing analytics and marketing agility of firms: The impact of the decision-making role. *J. Mark. Anal.* **2024**. [CrossRef]
46. Smith, A.A.; Li, R.; Tse, Z.T.H. Reshaping healthcare with wearable biosensors. *Sci. Rep.* **2023**, *13*, 4998. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.