

Challenges and best practices for digital unstructured data enrichment in health research: a systematic narrative review

Jana Sedlakova^{1,2,3}, Paola Daniore^{1,2}, Andrea Horn Wintsch^{1,5,6}, Markus Wolf^{1,7}, Mina Stanikic^{1,2,4}, Christina Haag^{1,2,4}, Chloé Sieber^{1,2,4}, Gerold Schneider^{1,8}, Kaspar Staub^{1,9}, Dominik Alois Ettlin^{1, 10}, Oliver Grübner^{1,11}, Fabio Rinaldi^{1,12,13,14,15}, Viktor von Wyl^{1,2,4} for the University of Zurich Digital Society Initiative (UZH-DSI) Health Community

Affiliations:¹ Digital Society Initiative, University of Zurich, Zurich, Switzerland; ² Institute for Implementation Science in Health Care, University of Zurich, Zurich, Switzerland; ³ Institute of Biomedical Ethics and History of Medicine, University of Zurich, Zurich, Switzerland; ⁴ Epidemiology, Biostatistics and Prevention Institute, University of Zurich, Zurich, Switzerland; ⁵ Center for Gerontology, University of Zurich, Zurich, Switzerland; ⁶ CoupleSense: Health and Interpersonal Emotion Regulation Group, University Research Priority Program (URPP) Dynamics of Healthy Aging, University of Zurich, Zurich, Switzerland; ⁷ Department of Psychology, University of Zurich, Zurich, Switzerland; ⁸ Department of Computational Linguistics, University of Zurich, Zurich, Switzerland; ⁹ Institute of Evolutionary Medicine, University of Zurich, Zurich, Switzerland; ¹⁰ Center of Dental Medicine, University of Zurich, Zurich, Switzerland; ¹¹ Department of Geography, University of Zurich, Zurich, Switzerland; ¹² Dalle Molle Institute for Artificial Intelligence (IDSIA), Switzerland; ¹³ Department of Quantitative Biomedicine, University of Zurich, Zurich, Switzerland; ¹⁴ Fondazione Bruno Kessler, Trento, Italy, ¹⁵ Swiss Institute of Bioinformatics, Switzerland

Corresponding Author: Prof Viktor von Wyl, Institute for Implementation Science in Health Care, University of Zurich, 8006 Zurich, Switzerland. Email: viktor.vonwyl@uzh.ch

Funding: This study was partially funded by the Digital Society Initiative (DSI).

Date: July 8, 2022

Word Count: 9699

Abstract

Digital data play an increasingly important role in advancing medical research and care. However, most digital data in healthcare are in an unstructured and often not readily accessible format for research. Specifically, unstructured data are available in a non-standardized format and require substantial preprocessing and feature extraction to translate them to meaningful insights. This might hinder their potential to advance health research, prevention, and patient care delivery, as these processes are resource intensive and connected with unresolved challenges. These challenges might prevent enrichment of structured evidence bases with relevant unstructured data, which we refer to as digital unstructured data enrichment. While prevalent challenges associated with unstructured data in health research are widely reported across literature, a comprehensive interdisciplinary summary of such challenges and possible solutions to facilitate their use in combination with existing data sources is missing.

In this study, we report findings from a systematic narrative review on the seven most prevalent challenge areas connected with the digital unstructured data enrichment in the fields of cardiology, neurology and mental health along with possible solutions to address these challenges. Building on these findings, we compiled a checklist following the standard data flow in a research study to contribute to the limited available systematic guidance on digital unstructured data enrichment. This proposed checklist offers support in early planning and feasibility assessments for health research combining unstructured data with existing data sources. Finally, the sparsity and heterogeneity of unstructured data enrichment methods in our review call for a more systematic reporting of such methods to achieve greater reproducibility.

Introduction

Digitalization has given access to a broad variety of digital unstructured data that contain health-relevant information and can substantially contribute to health research. Digital data in healthcare originate from a wide array of sources, ranging from structured clinical data, such as laboratory test results or patient-reported outcome measures, to unstructured data, such as free text data, collected within or outside of a clinical setting.¹ This wealth of data holds great potential to advance health research, prevention, and patient care delivery. However, over 80% of digital health data is available as unstructured data,¹ requiring new forms of data processing and standardizing that prove challenging to health researchers. The challenging nature of digital unstructured data is also reflected in the fact that these data are often not specifically collected for research purposes (e.g., data from social media).

Unstructured data are commonly defined as data that are not readily available in predefined structured formats such as tabular formats.^{2,15,21,27} However, there is no unified, standardized definition of unstructured data in health research. In the literature, unstructured data are often referred interchangeably as “big data”, “digital data”, “unstructured textual data” and described as “high-dimensional”, “large-scale”, “rich”, “multivariate” or “raw”.^{1,3,21,25,26,28}

Unstructured data can be utilized on their own or be combined with other data sources to enable *data enrichment* in health research. In this context, we refer to digital unstructured data enrichment to describe the process of augmenting the available evidence base in health research, which mostly consists of structured data with unstructured data.⁴ For example, open-ended patient self-reports or smartphone data can be used to complement longitudinal laboratory, clinical, and survey data.^{20,30,42} Through digital unstructured data enrichment, further insights into individuals’ lifestyles and behaviors can be gained due to the real-time measurements and monitoring data in a natural living environment, contributing to digital phenotyping⁵ and better understanding of health risks or diseases.³⁰ Furthermore, it can enable one to access under-researched population groups (e.g., ethnic minorities)⁶ and to gain a deeper understanding of participants’ daily life contexts over longer time periods, as well as outside of clinical settings.¹⁸ As such, this wealth of integrated data can foster

personalized, adaptive, and just in time health status assessments that can be of greater relevance to the study participants.⁷

While the abundance of digital unstructured data presents opportunities in advancing health research, methodological challenges surrounding their extensive preprocessing requirements for meaningful information extraction and integration persist.^{8,15,19,21,30} These challenges are accentuated as digital unstructured data are increasingly used to develop AI/ML models on unsupervised approaches, rather than on the standard supervised approaches.⁹ As a result, the established scientific process of creating and testing hypotheses is challenged in such a way that hypotheses are more strongly linked with the available data themselves.¹⁰ These persisting challenges and methodological developments are currently not addressed in the literature, as available methods mainly inform the pre-processing or optimization of computational possibilities with digital unstructured data, rather than informing health research study planning and conduct. As such, there is a need for guidance based on standards and best practices integrating different disciplines to inform the initial phases of study planning in health research with digital unstructured data.

Aims

This systematic narrative review aims to explore current standards and requirements to use digital unstructured data and their combination with existing data in health research. Specifically, we aim to answer the following research question:

How can health researchers enable the proper (systematic, reliable, valid, effective, and ethical) use of digital unstructured data to enrich a knowledge base from available data sources?

To answer this research question, this review 1) identifies and describes the main challenge areas associated with the use of unstructured data to enable digital unstructured data enrichment in health research; 2) provides a summary of possible solutions for common challenges associated with digital unstructured data enrichment; 3) provides guidance for the initial assessment of whether the inclusion of unstructured data is a feasible and appropriate for the study intended research tasks.

The goal of this review is to inform the planning and implementation surrounding the use of unstructured data in health research to enable knowledge enrichment from a methodological perspective.

Methodology

Definitions

We define *unstructured data* in accordance with the literature as raw data that are not in a pre-defined structure (e.g., tables) or data that may be structured, but still require substantial pre-processing or feature extraction effort.^{15,19,21,30} Furthermore, we define *digital unstructured data enrichment* as the use of unstructured data in combination with other data sources to contribute to relevant domain knowledge in health research and clinical practice.

In this review, we consider text data, electronic health records (EHR), sensory data from wearables and other devices, including electroencephalogram (EEG) as common sources of unstructured data. Despite their widespread use in health research, we did not consider imaging data in this review, as these data are often bound to manufacturer-proprietary algorithms, creating specific challenges in the enrichment process that may not generalize to other unstructured data types.

Search Strategy

We conducted a systematic narrative review guided by the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) 2020 statement.¹¹ Our study selection was guided by the inclusion and exclusion criteria displayed in **Textbox 1** and **Textbox 2**, respectively. We performed our search on PubMed and PsycInfo for 1) general overview articles, 2) primary research articles, and 3) articles describing databases, all including relevant information on digital unstructured data enrichment. Our search was restricted to articles from the fields of neurology, cardiology, and mental health. These were chosen due to the high prevalence of unstructured data availability in these fields and their established use for research and healthcare.^{12,13} The complete search syntax including all keywords can be found in **Appendix 1**.

Screening was conducted in two phases. In the first step, we screened the titles and abstracts from the studies based on the inclusion criteria (**Textbox 1**). In the second step, we performed a full-text screening of the articles selected in the first step and excluded articles that matched the criteria outlined in **Textbox 2**. In both steps, one investigator (JS) assessed all articles and a second investigator (PD) performed checks on a randomly selected sample of articles for each screening phase. Any disagreements were discussed and, if required, a decision was achieved through the principal investigator (VvW).

Textbox 1. Literature Review Inclusion Criteria

1. Published, peer-reviewed articles from 2016-2021.
2. Articles written in English.
3. Articles from the field of neurology, cardiology, mental health, or focusing on one of the diseases listed in the keywords.
4. Articles mentioning various sources of unstructured data and structured data in one of the three defined health fields.
5. Articles discussing challenges, limitations, or gaps of the integration of unstructured data in health research.

Textbox 2. Literature Review Exclusion Criteria

1. Articles focusing on imaging analysis or bioinformatics.
2. Articles outside of the three health-research areas: cardiology, neurology, mental health.
3. Articles including only structured data.
4. Articles leveraging a single data source (that is, no data enrichment).
5. Articles not addressing issues linked to the integration of unstructured data (or structured data with unstructured data).
6. Articles without empirical basis that are only *mentioning* integration of unstructured without further analysis, discussion, or explanations.

7. Systematic Reviews (narrative or literature reviews are included).

8. Protocols.

Data Extraction and Synthesis

Data extraction was standardized yet developed iteratively. The initial data extraction was based on standard study characteristics and guided by our research question. During the full-text screening, seven overarching topics related to digital unstructured data enrichment were identified and used for data extraction. The topics were the following: 1) medical field and subfield of the study, 2) main motivation for unstructured data integration, 3) data enrichment scope (e.g., gathering accurate information about disease severity), 4) type(s) of unstructured data, 5) limitations of unstructured data (e.g., quality/completeness), 6) challenges of data integration, and 7) proposed or discussed approaches for overcoming the mentioned challenges.

A narrative synthesis of the results was conducted to provide an overview on the challenges and proposed solutions related to the digital unstructured data enrichment. This choice was also influenced by the heterogeneity of included studies that ranged from overview papers to original research studies. To address study aims 1 and 2 (i.e., description of common challenges and their possible solutions), the extracted study data on the topics 5 and 6 (i.e., limitations and challenges associated with enabling digital unstructured data enrichment) were grouped into challenge areas. These challenge areas were defined on the basis of major overarching topics 5 and 6 identified after a first full-text screen of the included studies. The challenge areas include not only topics directly connected with data enrichment, but also related to the unstructured data use itself, as this is an essential requirement to enable digital unstructured data enrichment. For each challenge area, relevant possible solutions to tackle the challenges were summarized. For study aim 3 (i.e., providing guidance), we developed a preliminary checklist based on findings from our literature review to guide early study planning and feasibility assessment steps for studies aiming to include unstructured data. To this end, the identified challenge areas from study aim 1 were re-phrased into checklist questions and ordered according to the common

study planning stages in health research.¹⁴ Finally, the checklist was complemented and refined based on domain-specific expertise represented by the interdisciplinary team.

Results

Our database search yielded 358 articles (**Figure 1**). Overall, 28 articles were included for assessment in this review.

General description of included studies

A description of the 28 included articles^{15,16,17,18,19,20,21,22,23,24,25,26,27,28,29,30,31,32,33,34,35,36,37,38,39,40,41,42} is presented in **Supplementary Table 1 (Appendix 2)**. The most frequently discussed types of unstructured data sources in the selected articles were electronic health records (n=11) and sensor data (n=7). The most commonly cited motivations for digital unstructured data enrichment were to include of more objective measures in their research, for example, to improve understanding of disease mechanisms and disease prediction, and to strengthen the existing evidence base in precision medicine, real-time monitoring, and real-world data collection.

The most prevalent challenge areas in enabling digital unstructured data enrichment were: 1) the lack of meta-information for unstructured data (n= 6), 2) standardization issues (n= 20) 3) data quality and bias in data (n= 12), 4) infrastructure and human resources (n= 12), 5) finding suitable analysis tools, methods and techniques (n= 14), 6) alignment of unstructured data with a research question and design (n= 11), as well as 7) legal and ethical issues (n= 11). These challenges span across all study stages involving data in a health research study: from data collection to data interpretation. Definitions of the main challenge areas and a brief explanation of their relevance for health research are given in **Supplementary Table 2 (Appendix 2)**.

Challenge Areas

In the next sections, we summarize the seven identified challenge areas associated with enabling digital unstructured data enrichment in health research and the proposed possible solutions to address them.

1. Lack of meta-information for unstructured data

CHALLENGES: Lack of meta-information (e.g., describing data structure and properties or sample population) has been acknowledged as an obstacle for unstructured data findability, integration, interchangeability, and interpretation.^{15,19,26,27} Insufficient meta-information might limit the translation of a study's findings into clinical practice^{15,26} as important contextual information, such as information on the time in which the data were collected might be missing to assess the usability and correct interpretation of data.¹⁵

POSSIBLE SOLUTIONS: Proposed possible solutions included the standardization of meta-information (e.g., through a standardized format for meta-information through open science standards), which may also resolve issues of data interpretation and their alignment with research questions and designs.^{15,19,21,27,35} Specifically, one suggestion was to provide information for four important aspects in each study: subjects included in the data, context of collection, observations, and time of data collection.¹⁵ Moreover, a greater availability of standardized meta-information was suggested, as this would facilitate to determine the suitability of specific unstructured data for a given research question.^{15,21}

2. Standardization issues

CHALLENGES: The most frequently discussed challenge (n= 20) was the lack of a standardized framework for the description of disease phenotypes (e.g., symptoms, clinical presentation), as well as a lack of commonly defined terminologies, ontologies, and data labels.^{15,16,17,18,19,21,22,25,26,27,28,29,35,37,40,42} For example, different terms may be used for a seizure with alteration of consciousness by different physicians¹⁶ or for the administration of a specific dose of a given drug.²² These issues are particularly prevalent in EHRs or clinical annotations¹⁹ where, for example, terminologies and phenotyping may differ across healthcare settings or change over time.^{16,21,23} There is also an observed lack of standardized data management methods^{27,37} and regulatory standards to guide and assess the use of novel technology and their associated unstructured data in clinical applications such as clinical trials.^{15,30}

POSSIBLE SOLUTIONS: In most articles, harmonization of data formats, data models, terminologies, ontologies, and analytical tools, as well as working practices were proposed as possible solutions to standardization issues.^{15,20,21,22,24,27,28,29} A consensus of standards across the entire data flow,^{15,21,22,27,28} the effective use of datasets,²⁷ data optimization,²⁹ data consistency^{23,41} and replicability of the studies^{20,21} were also suggested as a means to foster data sharing. The adoption of unified data standards was considered to be important in both academic and industry settings.²⁸

To improve standardization efforts and data sharing, the systematic adoption of FAIR (Findable, Accessible, Interoperable, Reusable) Guiding Principles for scientific data management⁴³ was proposed.²⁸ Other authors mentioned the need for specialized organizations to promote harmonization of terminologies in health research.^{21,28} An example is the consortium behind the Fast Health Interoperability Resources (FIHR) standard to enable “interoperable communication and information sharing between various healthcare systems”.²⁸

3. *Data quality and biases in data*

CHALLENGES: Data quality of unstructured data was frequently cited as an important challenge for evidence creation.^{17,19,20,21,23,26,27,29,32} Unstructured data are often collected for purposes other than research and may lack systematic collection and the rigor of study-based measurements, thus often leading to missing data.^{19,20,21,23,26,27,29,39} In medical records, for example, missing data can occur because health care professionals may omit some information or because of patients’ refusal to share data.²⁶ The challenge of data quality is reinforced by data inconsistencies and inaccuracies.^{20,21,23,26,29,32} Other recurrent challenges stem from biases in data collection – mainly in the form of selection and information bias^{17,19,23,24,26,27,29} – and confounding.^{21,23} Selection bias was mentioned, for example, in the context of studies where the sample comprised only of individuals who have the digital literacy skills or interest to share unstructured data from social media or wearable sensors (5, 8). Information bias, such as observer bias, was often mentioned in the context of making errors with data in EHRs use and big data analytics.^{17,23,24,26} Further biases may establish themselves in analyses if processing algorithms were trained on biased data.^{19,23,24} Finally, the quality and continuity of data might be

negatively impacted by technical issues that can arise, for example, by software updates of wearable sensors.²⁷

POSSIBLE SOLUTIONS: Several strategies were proposed for assessing and ensuring data quality.^{15,19,21,23,26,27} For studies that use digital health technologies, one study cited a recommendation from the European Medicines Agency (EMA) urging researchers to define “small, well-defined, meaningful measures followed by a data-driven development path”.²⁷ Furthermore, possible data quality issues should be considered for all study phases, including preprocessing, feature extraction or analysis.¹⁹ First, preprocessing should yield only verified and valid dataset that properly combines the unstructured data with other data sources, for example by ensuring that study samples are representative of the populations that are being studied. Second, following feature extraction, data should be critically assessed for their validity and meaning. Finally, analytical methods should be aligned with the research goals of description, prediction, or prescription of the study in such a way that bias is reduced.

Other studies highlighted the need for a data quality standard checklist such as Data Access Quality and Curation for Observational Research Design (DAQCORD).^{19,44} This checklist should provide a priori guidance for planning of large-scale study data collection and pre-processing,⁴⁴ thereby countering the pervasive practice of post-hoc methods for data cleaning.¹⁹ Other proposed possible solutions included the use of meta-information to increase data quality, to detect potential biases in the data,¹⁵ to enable cross-referencing of multiple data sources involving the same individuals, as well as to encourage the comparison of results.²¹ Imputation procedures for addressing missing data, as well as algorithms for checking data quality were also recommended.¹⁹ Furthermore, the inclusion of study participant feedback can inform data collection and processing and improve the relevance of study findings for the intended target population.²⁷

4. Infrastructure and Human Resources

CHALLENGES: Several studies pointed out challenges related to infrastructure availability, including databases, or open-source platforms.^{15,19,21,22,25,26,27,28} Infrastructure challenges can be particularly

problematic when healthcare data are spread across multiple medical systems that lack connection or interoperability, thus creating isolated data clusters.^{22,25,28} Difficulties in data linkage can also emerge when information system architectures cannot accommodate data standardization and other linkage processing tools.¹⁹ Furthermore, the lack of skills and formal training opportunities for infrastructure utilization or inadequate knowledge of novel statistical tools and methods for combining unstructured data with other data sources can inhibit their use in health research.^{21,26,27}

POSSIBLE SOLUTIONS: Improvements such as searchable catalogues, databases and the availability of open platforms can mitigate infrastructure-related challenges.^{16,18,19,20,21,24,27,28} Similarly, the availability of infrastructure for the storage and integration of unstructured data can enable collaborative efforts, facilitate standardization, and foster the alignment of unstructured data with good research question development and research design.²¹ Furthermore, the availability of secure collaborative platforms and repositories for data sharing through open science can enable independent knowledge gain and foster new research studies.^{16,27} Meta-databases or catalogues that facilitate the discovery of open data and linking data across public repositories can also facilitate digital unstructured data enrichment.^{21,28} Several studies further suggested that platforms for integrating datasets from various sources should have a modular, flexible, and scalable structure^{16,19,28} and recommended to define the purpose and goals of such platforms during their development.^{20,24} Open data and open software repositories also provide more opportunities for external validation of novel algorithms or (electronic) clinical outcome measures.¹⁸ Finally, awareness about novel digital unstructured data enrichment methods, their methodological requirements, and the need for specialized training opportunities should be increased.^{20,21}

5. Finding suitable analysis tools, methods, and techniques

CHALLENGES: The complexity of analyses and appropriate methodological choices associated with the unstructured data enrichment in health research are challenges that were addressed in multiple studies.^{16,19,21,23,24,26,28} Typical features of unstructured data such as high volume or complexity may be overwhelming for researchers due to a lack of methodological knowledge.²⁶ Furthermore, the validity

of results may be decreased by algorithms that are either not trained sufficiently or may need recurrent fine-tuning to ensure that they create a model representative of its intended purpose and without biases.^{18,19,26} Working with unstructured data requires specific expertise, typically from data scientists. However, the lack of supply of data scientists or the failure to build effective collaborations with external experts were also cited as impediments to managing the complexity of unstructured data.^{21,24,28} Furthermore, there is a lack of guidelines and standards to guide decisions on which tools, methods, and analytical approaches to use when using unstructured data in health research.^{21,26}

We further observed a discrepancy in approaches to reduce the complexity of unstructured data (e.g., using feature extraction) in our studies. While some authors argued that complexity reduction is a feasible and appropriate method to enhance unstructured data integration, others voiced concerns that complexity reduction can also reduce richness of unstructured data – particularly in the context of EHRs.^{16,42}

POSSIBLE SOLUTIONS: The complexity of unstructured data calls for increased collaboration among different experts. The increasing need for interdisciplinary efforts among health researchers, data scientists, biostatisticians, and health-care professionals was highlighted by most sources.^{15,16,17,18,20,21,27,29} Some authors emphasized the need for a novel profession that combines expertise in health research and informatics.^{21,29} Many also called for greater attention to trainings of health researchers regarding novel methods for using and combining unstructured data with other data sources.^{18,20,21} The need for specific sets of skills, resources, and guidelines for the successful implementation of big data tools into clinical workflows was further mentioned as a requirement to manage unstructured data complexity.²³ Furthermore, some authors called for more efforts to develop and establish validated algorithms to process and integrate data.²⁶ One suggestion was to “provide AI with more ‘functional’ information, such as domain-specific medical reasoning processes and policies based on heuristic-driven search methods derived from human diagnostician methods”.²⁸ In the field of mental health, it was advised to complement data-driven research with qualitative research to strengthen the relevance and meaning of results.²⁴

6. *Alignment with a research design and/or research question*

CHALLENGES: The difficulty of finding suitable datasets and their subsequent, critical evaluation for clinical relevance was discussed from several perspectives.^{17,18,21,23,24,26,29} One study strongly warned against adjusting the research agenda to the data that are available.²⁴ Furthermore, the fact that unstructured data or technologies generating these data were not designed for scientific purposes^{17,18,20,24} might lead to misinterpretation of the data.²⁴ The lack of contextual (meta) information, for example about the data generation process, and observational nature of many sources of unstructured data may limit the value of the data for their use in robust, replicable confirmatory analyses (e.g., regarding disease etiology or intervention).^{17,24} The need for further and robust validation of results or outcomes from unstructured data analyses was a further topic of concern.^{20,23,24} For example, predictive models need further validation before being integrated into clinical settings²³ and informing clinical decision-making.²⁶ Similarly, while linked EHRs are suitable for generating research questions, unstructured data should not be used for influencing clinical practice without prior validation.²¹

POSSIBLE SOLUTIONS: It should be ensured that unstructured data are relevant for a research question and desired therapeutic effect.¹⁸ When working with data from digital health technologies, the EMA recommendation framework - that was developed with the collaboration with industry representatives with the aim to provide insights and guidance on validation and qualification processes of digital technologies⁴⁵ - could be consulted for guidance with research question design.²⁷ Another recommendation was to align large-scale research projects using unstructured data with clinical priorities and outcome-focused research.²⁰ Similarly, the choice of analytical tools depends on the goals of health research: description, prediction, or prescription.¹⁹ Thus, setting clear research goals might help with the choice of appropriate analytical tools and methods. Finally, unstructured data should be used rather with complementary and enrichment purposes than as a replacement of other traditional methods or datasets.^{18,23,24,30}

7. *Ethics & Legal Issues*

CHALLENGES: The most frequently mentioned ethical challenges concerned privacy protection, informed consent and preservation of individual agency over data use.^{19,20,21,23,25,26,27,32} Further challenges connected with digital unstructured data enrichment include inappropriate patient profiling²³ and decreased participants diversity due to low digital literacy skills reducing some participants' contributions to certain types of unstructured data (e.g., from social media use).²⁷ Furthermore, current deidentification and anonymization practices may still allow patient-linkage. This is, for example, enabled when a combination of data on unusual physical conditions of a patient from a local hospital or a combination of gender, age and admission date might be unique enough to identify a subject and connect it with consumer-level data.^{19,42}

POSSIBLE SOLUTIONS: Strategies for preserving data privacy and security were discussed in multiple studies.^{18,19,20,21,25,26,27,35} Some authors proposed to develop a new social contract and a broad consent model to balance the benefits of data usage and privacy concerns.^{20,21,26} Unified rules for data governance across fields and sectors might contribute to systematic privacy protection and confidentiality,¹⁹ such as through unified procedures for data anonymization. Additionally, the importance of engagement with regulatory agencies in early stages of research was emphasized to ensure alignment of unstructured data processing with best practices.²⁷ Finally, independent agencies or governing bodies were proposed to oversee and ensure safe data sharing, preservation of intellectual property and valid applications.^{18,19}

Additional Recommendations

During the literature review, we identified two additional, overarching recommendations for unstructured data use, which are described subsequently.

Collaborations with all stakeholders

Several sources stressed the importance of stakeholder collaboration in health research when combining different data sources for knowledge enrichment.^{20,27,28} The inclusion of with public and patient advocacy groups and other relevant stakeholders was highly recommended²⁰ to ensure wide public acceptance and patient trust.^{21,35} Broad stakeholder involvement was also seen as crucial to

increase data sharing and to minimize wasted efforts from research study duplication.²⁷ Collaborative efforts among academic and commercial organizations (e.g., digital device or sensor manufacturers) can facilitate large-scale data integration and create synergies.^{20,28} Stakeholder and patient engagement during in the unstructured data integration, analysis, and interpretation provides relevant context and feedback on the meaningfulness of results.^{20,27}

Documentation and transparency

Proper documentation and transparency during the entire data flow were repeatedly mentioned as essential steps to achieve reliability, replicability, reproducibility and validity of studies, as well as facilitating the standardization efforts.^{17,24,27,28} The EMA framework emphasizes documentation as an important means to achieve reliability, repeatability, accuracy, clinical validity, generalizability, and clinical applicability of the novel methodologies.²⁷ In the context of digital health technologies, United States Food and Drug Administration (FDA) recommendations suggest documenting the device and algorithm input and output, and to provide plans for data loss minimization, missing data handling, or patient inclusion for results. Furthermore, the FDA recommendations call for transparency of all processing steps from raw data to algorithm and at all data workflow stages.²⁷ Transparency regarding the analysis process can also assist with the assessment of whether study findings were clinically significant.²⁴ Specifically, studies relying on large databases will produce many statistically significant, but clinically meaningless results. This “overpowering” of statistical tests by large sample sizes should be made transparent through reporting of effect size determinants and complementation by clinical interpretation.¹⁷

Proposal for a feasibility and planning checklist for unstructured data enrichment

Many studies highlighted the need for further research and guideline development on best practices to use and integrate unstructured data in health research.^{18,20,21,22,23,28,29} In **Table 1**, we provide a set of guiding questions to inform early study planning and the assessment of the feasibility of studies.

These questions are based on the described challenge areas, which have been expanded to align with the breadth of proposed solutions from our review.

Table 1. The checklist for early study planning and the assessment of the feasibility of studies using digital unstructured data

Key Issues	Comments
Sufficient Metadata & Documentation for Unstructured Data ^{15,19,21,27,35}	Meta-information can describe primary data and provide contextual information about data collection, pre-processing, or interpretation. Meta-information is especially important for data that were collected for purposes other than research or data from wearables and other electronic devices.
1. Is meta-information for the unstructured database available and where?	Meta-information should be findable and well documented.
2. Can meta-information offer sufficient contextual information for data interpretation?	Meta-information should include: ¹⁵ <ol style="list-style-type: none"> 1. Person: e.g., subject ID, medical history, or demographics 2. Context of collection: environment, study ID, or procedure description 3. Observations: e.g., technology-affiliated site location, technology type, or notes made by an observer (e.g., a clinician) 4. Time of data collection: e.g., time source, time zone, or medication schedule
Standardization Options for Unstructured Data ^{15,16,17,18,19,21,22,25,26,27,28,29,35,37,40,42}	

3. Are data transformable into a standardized format?	A standardized format can be a tabular format. Many different standards for clinical data already exist. For example, for EHRs, the Fast Healthcare Interoperability Resources might be useful. ⁴⁶
4. Do data already contain standardized syntax/terminology or can such standards be applied?	Clinical information such as terminology and coding for diseases (e.g., ICD-10) might differ across databases. For data integration, sharing and reproducibility, standardized syntax and terminology are important.
5. Does the dataset contain standardized semantics/ontology, or can such standards be applied?	Standardized ontology describes logical relations between core concepts to structure the description of data and foster interchangeability and consistency of data. For example, Systematized Nomenclature of Medicine—Clinical Terms (SNOMED CT) is a comprehensive medical terminology used for electronic health data. ⁴⁷
Data Quality ^{19,23,41}	For observational studies, the checklist DAQCOR ⁴⁴ might be a useful starting point.
6. Can the consistency of data be secured? Are strategies/methods/steps available and included in the data management to secure consistency of data?	Consistency of data refers to the concept that the same data stored in separate places or separate time points still match, meaning contradictory conclusions cannot be derived from the given data. ⁴⁸ For example, can be ensured that archived/backed-up/repository-deposited information can be kept up to date?
7. Is the dataset without a significant amount of missing data?	Due to selective reporting, EHRs may lack important data because clinicians did not

	deem them relevant, or patients did not want to share them. ²⁶
8. Are strategies/methods available and/or defined for dealing with missing data?	Data may be missing for different reasons. They may be ‘missing at random’, they may be missing because the information was deemed irrelevant (e. g., not collected or not relevant for research question), because of branching or procedural logics (e.g., data are only collected under certain conditions). This knowledge also informs the feasibility of multiple imputation techniques, which assume at least some randomness (either systematic or non-systematic) in missing data.
Data Validity ^{15,18,19,26}	
9. Is the population for which the data are available representative of the target population?	Data might be only available for a limited population, for example., only a particular population group used the device collecting the data. Also, physicians’ notes are only available for a select subgroup (e.g., persons with a more severe clinical presentation). A sound understanding of the data generation process (possibly informed by meta-information) is essential.
10. Are strategies implemented to prevent or minimize the risk that the data are affected by selection or information biases?	Biases can prevent that the measures or outcomes correspond to their true value. Epidemiological and medical research commonly distinguishes between three types of biases: ⁴⁹ Selection bias occurs when the selection of study participants alters the exposure-outcome relationship (not to be confused with external

	<p>validity/representativeness). Information bias occurs when the ways of how data is collected impair data accuracy. Confounding refers to an observed relationship between exposure and outcome, which is influenced by a third, unaccounted variable (e.g., lung cancer is more prevalent among persons who drink alcohol, but smoking is also associated with alcohol consumption).⁵⁰ Note that terminologies regarding biases may differ across scientific disciplines.</p>
<p>Alignment with Research Question and Design^{18,23,24,27,30}</p>	
<p>11. What is the purpose or motivation of enriching a dataset with unstructured data?</p>	<p>For example, unstructured data can provide additional insights into individuals' lived experiences or provide information in higher temporal resolution than standard data collection approaches (e.g., surveys).</p>
<p>12. Can the purpose of unstructured data be linked with a well-defined research question?</p>	<p>It is advisable to specify clearly defined and operationalized hypotheses before conceptualizing and conducting the study. Considerations are needed whether unstructured data enrichment increases the chances for successfully testing of these pre-specified hypotheses.</p>
<p>13. Can the use of unstructured data be aligned with the planned research task (description, prediction, exploration, explanation, application)?</p>	<p>The aim of unstructured data integration might differ depending on research tasks, e.g., description (such as describing a disease progression), prediction (predicting outcomes), exploration (to find new patterns or generate a new hypothesis),</p>

	<p>explanation (to establish causality) or application (such as the development of a practical tool for diagnosis).</p>
<p>14. Can the combined dataset lead to relevant, novel insights?</p>	<p>It should be considered what added value can be expected by the enrichment with unstructured data. Examples are deepened qualitative or quantitative insights, more real-time time data, or a stronger participant-centeredness.</p>
<p>Infrastructure for Processing and Analysis^{15,16,17,18,20,21,23,27,28,29}</p>	
<p>15. Does the research team have relevant skills/ or access to experts to approach for integration of unstructured data?</p>	<p>Interdisciplinary teams should include persons with strong (quantitative and/or qualitative) research methods skills as well as subject domain knowledge (e.g., specialists in a particular clinical area).</p>
<p>16. Can the interdisciplinary work be well established?</p>	<p>Define strategies to include persons with the necessary skills in the project teams, e.g., through existing networks, through consulting services (e.g., statisticians), referral by colleagues.</p>
<p>17. Can any duplication of research be excluded?</p>	<p>It is advisable to search for and summarize existing literature. It might be useful to check open data sources, platforms, database aggregators or searchable catalogues</p>
<p>Availability of suitable analysis tools, methods and techniques ^{18,20,21, 24,26,28}</p>	
<p>18. Are the appropriate methods, tools, and techniques available?</p>	<p>The integration of unstructured data with other data sources requires a set of</p>

	different methods, tools and techniques from informatics, data science, software development and others. The complexity of data opens many possibilities for analysis and statistical methods and their choice should be well justified.
19. Were the analysis methods, tools, and techniques chosen in a way that does not increase the risk of biases?	Analytical methods should ideally match the pre-specified study questions and hypotheses – not the other way around.
20. Can unstructured data be structured without significant loss of richness or other limitations? If no, will such limitations be reported/documentated?	For example, qualitative information about patient experience from EHRs might get lost. Consider the integration of qualitative information as part of the analysis. Consider spot-checks and validation of quantitative findings using unstructured data (e.g., through random chart reviews).
Expected quality of evidence of combined database ^{17,19,20,21,23,26,27,29,32}	
21. Can the methodology of hypothesis testing be well defined?	Given a set of pre-specified hypotheses: Are the data and planned methods suitable to detect the effect of interest (as indicated by, e.g., an a priori power analysis)?
22. Can the results be sufficiently validated to serve as research evidence?	Validation means the testing of (prediction) models and study findings in other, previously unused data. Validation pertains to testing whether the study findings can be applied to other similar individuals outside of the study and whether a statistical relationship, for example between cause and effect, can be generalized. ⁵¹

<p>23. Can unstructured data be technically combined/merged with structured data?</p>	<p>What are the links between structured and unstructured data? Are there shared unique identifiers in both databases? Do the unstructured data need matching by specific time-points? Common challenges are that structured and unstructured data are not collected at synchronized time points and/or for all participants.</p>
<p>24. Can input from patient/population from whom the data was collected be meaningfully included in the study?</p>	<p>Including qualitative data of individuals' input about their experience, for example, with wearable sensors can provide important contextual information for ensuring quality and relevance of data and the study.</p>
<p>Ethical and Legal Aspects^{18,19,20,21,25,26,27,35}</p>	
<p>25. Have ethical and data security requirements been clarified and reviewed? If not, is it planned to contact relative authorities and regulators be contacted to clarify privacy and ethical requirements?</p>	<p>Many studies involving health data require approval by ethics committees. Moreover, it may be advisable to seek contact with data protection officers upfront to assess and identify potential data security and privacy risks.</p>
<p>26. Have strategies for securing data privacy and security been clarified and are ready for implementation?</p>	<p>Linkage, processing, and analysis of unstructured and structured data require planning and consideration of the complete data life cycle. A data management plan should be put in place to outline rules and principles for handling data.</p>
<p>27. Can the data be fully anonymized or pseudonymized?</p>	<p>Deidentification efforts might not lead to full anonymization because an individual might be uniquely identified due to a specific piece or aggregates of information.</p>

Transparency, Reporting ^{17,24,27,28}	
28. Can all the steps of data collection and management be well documented?	With the steps of data management, we mean data collection, data source, data storage, data retrieval, data preprocessing, data analysis, data interpretation. Ideally, these considerations should be included in a data management plan and cover the full data life cycle.
29. Can be ensured that the documentation contain elements of established reliability, accuracy and validity of the studies?	<p>Reliability refers to the stability of findings⁵⁵</p> <p>Accuracy is the proximity of measurement results to the "true" value.</p> <p>Validity is the truthfulness of findings; the results of an experiment do measure the concept being tested.^{52,53}</p>
30. Is the analysis process being documented in a detailed fashion (e.g., inclusion of the description of analysis on the level of coding or data) that allows sharing and replication?	This is a requirement for open science. A study protocol can be a good starting point for documentation. Moreover, all preprocessing and analysis steps should be programmed/coded and commented on.
31. Can the documentation support generalizability and replicability of studies?	<p>Replicability is obtaining consistent results across studies aimed at answering the same scientific question.⁵⁴</p> <p>Generalizability means that the study results or outcomes are applicable also in other study settings or samples.⁵⁵</p>
32. Can limitations of data be reported?	For example, it should be reported whether unstructured data was only collected

	from a limited group of population or the collected textual data from social media was limited to long posts which might lead to a collection of data from a population group with specific characteristics.
33. Can the relevant technical steps of the data integration be reported?	This includes strategies, definitions and techniques for combining structured with unstructured data. For example, were data linked by person and for specific time points? What were assumptions and definitions used in the linkage process (e.g., was there a pre-specified time window within which two data points/assessments were considered as simultaneous)?
Reproducibility of pre-processing, feature extraction and analysis, Open Science ^{16,18,19,20,21,24,26,27}	Reproducibility means obtaining consistent results using the same input data; computational steps, methods, and code; and conditions of analysis. Reproducibility is closely connected with transparency, sufficient reporting and availability of data and methods. ⁵⁴ For reproducibility purposes and data sharing, FAIR principles might be particularly useful. FAIR ⁴³ means “Findability, Accessibility, Interoperability, and Reuse of digital assets”.
34. Can it be ensured that the preprocessing steps reproducibly yield valid intermediary/analytical data?	See points above regarding documentation (questions 28-33).

<p>35. Can it be ensured that the feature extraction algorithms reproducibly yield data for meaningful statistical processing?</p> <p>Are extracted features sensitive to changes/adaptations in algorithms?</p>	<p>See points above regarding documentation. Documenting algorithms may be challenging when relying on proprietary software.</p>
<p>36. Can raw data / intermediary data / analytical data be made openly available?</p>	<p>Whether and how data can be made available depends, for example, on data ownership, the availability of informed consents by participants, privacy risks, risks for re-identification.</p>
<p>37. Can raw data / intermediary data / analytical data be integrated into a well-designed open platform repository?</p>	<p>Many open data repositories with different requirements regarding data format or documentation exist. Many scientific journals demand a mandatory upload of certain data types into public repositories (e.g., for genetic sequence data).</p>

Discussion

Summary of findings

Our systematic narrative review provides an overview of challenges and best practices associated with the combination of unstructured data with other data sources in health research, which we refer to as digital unstructured data enrichment. In our review, we identified seven prevalent challenge areas in enabling digital unstructured data enrichment: 1) the lack of meta-information for unstructured data, 2) standardization issues, 3) data quality and bias in unstructured data, 4) infrastructure and human resources, 5) finding suitable analysis tools, methods and techniques, 6) alignment of unstructured data with a research question and design, as well as 7) legal and ethical issues. For each challenge area, we summarized proposed possible solutions together with two additional recommendations that span across all challenge areas. We also summarized literature and experience-based checklist questions to inform initial study planning about the feasibility of research studies aiming to complement existing health data with digital unstructured data.

Description of main requirements and solutions to enable unstructured digital data enrichment

All our studies revealed challenges associated with the digital unstructured data enrichment in health research, many of which might endanger scientific rigor and quality of health studies. For example, the frequently unclear suitability of digital unstructured data to address concrete research questions or allow for proper research study design^{18,21,26} may lead to possible biases, threatening the external and internal validity of studies. The validity of studies might also be endangered by applying not suitable analytical tools and methods. Furthermore, the findings of the study may lack generalizability limiting its use to specific research tasks and questions (e.g., hypothesis-generation).^{17,18,21,26} The lack of meta-information might hinder a proper interpretation of the data and consequently limit their use for enrichment purposes. Further problems are that the data can be placed so centrally that any bias will be strongly reflected in the results, also in the formation of data-driven categories. The most discussed challenge of standardization issues might hinder replicability and generalizability of research studies. Finally, ethical and legal issues, such as the risk of patient re-identification when disparate data sources are combined, pose additional challenges to digital unstructured data enrichment.

While many of the challenges to enable digital unstructured data enrichment are not specific to the use of unstructured data and are well known (e.g., data quality or standardization issues), other challenges, such as difficulties to align data with research questions or challenges pertaining to special skills or infrastructure needs, may be aggravated with the use of unstructured data due to their complexity. One of the key challenges might be the lack of open and collaborative platforms that can foster not only joint standardization but also validation efforts.^{15,21,28} Oftentimes, the attractive characteristics of unstructured data that might add value to research are the ones that pose the most challenges. The data granularity and large, often international, population-based sample can enhance disease understanding or monitoring but also lead to methodological challenges, for example, regarding validity and choice of tools for analyses.^{24,26}

The possible solutions and additional recommendations are important to sustain interchangeability, validity, reliability, generalizability, and reproducibility of studies. The review revealed that the possible solutions are less frequently and systematically discussed than the challenges. Several sources discussed challenges without referring to the existing solutions or offering new proposals for possible solutions. The complexity of the digital unstructured data enrichment is also reflected in the possible solutions that can address several challenge areas. The interdisciplinary collaboration, open science and transparency were one of the most requested possible solutions.

Requirement for guidance on digital unstructured data enrichment

Our review also revealed that despite the wide usage of unstructured data in health research and discussed challenges, there is a lack of a systematic approach and guidelines for researchers to address the observed challenges. Several of the selected articles acknowledged the need for more guidance,^{18,20,21,22,23,28,29} oversight or monitoring from agencies^{18,27,42} and interdisciplinary teamwork and exchange to establish methodological approaches in the context of utilizing unstructured data in health research.^{21,26} Only a few studies directly mentioned existing frameworks and standards such as

EMA recommendations,²⁷ FAIR principles,²⁸ openHR²¹ or DAQORD framework¹⁹ in this context. Recent efforts to provide guidelines are mainly focused either on a specific type of unstructured data or specific challenges, for example, guidelines and standards for the use of social media data,^{56,57} guidelines regarding the use of EHRs,^{58,59} checklists and frameworks for evaluating the measurements made by digital technologies⁶⁰ or algorithms used for data analysis.^{61,62} However, it is up for discussion whether these particular frameworks and guidelines are suitable to provide general guidance on challenges connected with digital unstructured data enrichment.

Our findings also reveal an underreporting of information relevant to digital unstructured data enrichment in health research. For example, current reporting guidelines such as STROBE⁶³ do not cover unstructured data-enriched analyses. In the assessed studies, challenges relevant to integrating unstructured data with other health data sources were rarely mentioned. Our studies usually provided description of data collection and preprocessing and addressed issues of noisy and missing data. However, they often lacked a description of data limitations or strategies how to ensure data quality. In light of the growing volume and importance of unstructured data in health research, experience sharing should be increasingly encouraged – either in published literature (e.g., also in appendices) or in other outlets. The lack of reporting and unavailability of guidelines not only hampers study reproducibility but also presents missed opportunities for learning and capacity building.

All this points to a growing need to define systematic ways of how to approach digital unstructured data enrichment in health research. The numerous challenges directly linked with unstructured data use or digital unstructured data enrichment should be reflected in a systemic guidance on how to properly integrate digital unstructured data in health research. Our review identified a special need for guidance to establish common standards to enable digital unstructured data enrichment to help researchers in the first stages of study planning and to assess the feasibility of studies integrating unstructured data.^{16,18,27} The checklist derived from our review provides a first, pragmatic step towards classifying challenges and developing methodologies in health research involving digital unstructured data enrichment. In next steps, we hope to encourage specific research fields to dive deeper into our

proposed checklist and adapt it to terminologies and issues that might be of a greater relevance in their respective research fields.

Limitations

Although based on a systematic search and extraction process, we restricted our search to a few specific research fields due to the prevalent and growing use of unstructured data in these fields. Furthermore, we did not include books and book chapters. Therefore, our overview is likely not comprehensive. Furthermore, we a priori excluded imaging data and bioinformatics data from our literature search, which are an important source of unstructured data, but are often analyzed with highly specialized tools. In the systematic narrative review, we did not specifically discuss challenges and obstacles that are linked with learning algorithms used for unstructured data integration or data analysis/interpretation. However, there is also precaution and guidance needed for choices about learning algorithms. Machine learning and deep learning algorithms are not immune to errors, biases and other limitations that can negatively impact validity, objectivity, and reproducibility of studies.

Conclusion

The integration of unstructured data into structured databases opens new avenues for more person-centered, contextualized, or more real-time analyses. However, multiple methodological and conceptual challenges demand attention, ideally even before an analysis is undertaken. A clear definition and focus on suitable study questions, interdisciplinary team-work, or transparent documentation and open science are key ingredients towards a more robust unstructured data enrichment methodology. Overall, our review also points to a need of more guidance – and possibly also standards for reporting results of digital unstructured data studies. Awareness should be raised among researchers to openly document encountered challenges and possible solutions in unstructured data enrichment projects to enable experience exchanges and learning. Moreover, existing reporting guidelines such as STROBE should consider adding specific instructions on the documentation of unstructured data enrichment processes.

Data Availability

The data from the papers that support the findings of this study are publicly available. All data used in this review can be provided upon request.

Author contributions

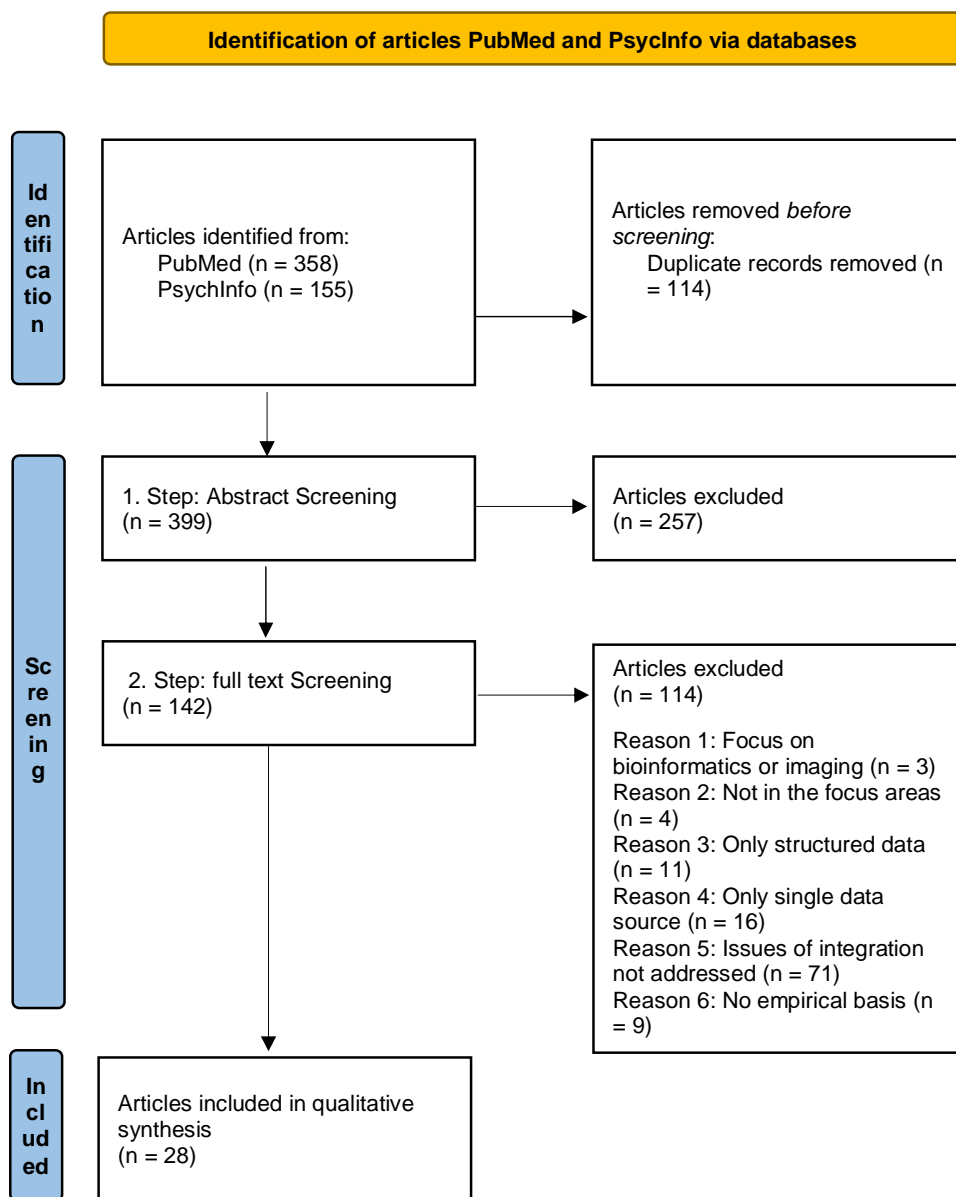
The review was conducted as a collaborative effort in the Health Community of Digital Society Initiative at the University Zurich (the members' list is to be found in the **Appendix 3**). The core research team (consisting of JS, PD, VvW) defined the initial plan for the review together with the research question and search syntax as well as conducted the literature search, screening steps, data extraction, data synthesis and wrote and revised the manuscript. MS, CH, CS assisted with the initial plan for the review, conducted the literature search, helped interpret the findings and revised the final manuscript. The entire process was discussed in the Health Community meetings. In addition, every step was consolidated, evaluated and approved by the Executive Committee (consisting of AHW, MW, GS, OG, DAE, KS) that provided further insights, comments and contribution to the manuscript.

Ethics Declaration

The authors declare no competing interests.

Correspondence should be addressed to Viktor von Wyl.

Figure 1. PRISMA Flow Diagram



References

- ¹ Kong, H.J. Managing Unstructured Big Data in Healthcare System. *Healthc Inform Res.* 25(1):1-2. (2019).
- ² Unstructured data. [Cited 7 July 2022]. Available from: <https://www.sciencedirect.com/topics/computer-science/unstructured-data>
- ³ Adnan, K., Akbar, R., Khor, S.W. & Ali, A.B.A. Role and Challenges of Unstructured Big Data in Healthcare. In: Sharma, N., Chakrabarti, A., Balas, V. (eds) *Data Management, Analytics and Innovation. Advances in Intelligent Systems and Computing*, vol 1042. Springer, Singapore. (2020).
- ⁴ Data Enrichment. [Cited 7 July 2022]. Available from: <https://www.sciencedirect.com/topics/computer-science/data-enrichment>
- ⁵ Huckvale, K., Venkatesh, S. & Christensen, H. Toward clinical digital phenotyping: a timely opportunity to consider purpose, quality, and safety. *NPJ Digit Med* 2:88. (2019).
- ⁶ Zhang, X., Pérez-Stable, E.J., Bourne, P.E., et al. Big Data Science: Opportunities and Challenges to Address Minority Health and Health Disparities in the 21st Century. *Ethn Dis.* 27(2):95-106. (2017).
- ⁷ Wang, Q.C. & Wang, Z.Y. Big Data and Atrial Fibrillation: Current Understanding and New Opportunities. *J Cardiovasc Transl Res.* 13(6):944-952. (2020).
- ⁸ Sheikhalishahi, S., Miotto, R., Dudley, J.T., Lavelli, A., Rinaldi F. & Osmani. V. Natural Language Processing of Clinical Notes on Chronic Diseases: Systematic Review. *JMIR Med Inform* 7(2):e12239. (2019).
- ⁹ Benedetto, F. & Tedeschi, A. Big Data Sentiment Analysis for Brand Monitoring in Social Media Streams by Cloud Computing. In: Witold Pedrycz and Shyi-Ming Chen (eds.), *Sentiment Analysis and Ontology Learning. Studies in Computational Intelligence* 639. Cham, Switzerland: Springer. p. 355. (2016).
- ¹⁰ Succi, S. & Coveney, P.V. Big data: the end of the scientific method? *Phil. Trans. R. Soc. A.* 377. (2019).
- ¹¹ Page, M.J., McKenzie, J.E., Bossuyt, P.M., Boutron, I., Hoffmann, T.C., Mulrow, C.D., et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ* 372:n71. (2021).
- ¹² Sim, I. Mobile Devices and Health. *N Engl J Med* 5;381(10):956-968. (2019).
- ¹³ Hulse, T. Challenges and solutions for big data in personalized healthcare. In: *Big Data in Psychiatry and Neurology*, ed. by Ahmed A. Moustafa, Academic Press, p 69-94. (2021).
- ¹⁴ Wissik, T. & Ďurčo, M. Research Data Workflows: From Research Data Lifecycle Models to Institutional Solutions. *Linköping Electronic Conference Proceedings* 123:8, 94-107. (2015).
- ¹⁵ Badawy, R., Hameed, F., Bataille, L., Little, M. A., Claes, K., Saria, S., Cedarbaum, J. M., Stephenson, D., Neville, J., Maetzler, W., Espay, A. J., Bloem, B. R., Simuni, T., & Karlin, D. R. Metadata Concepts for Advancing the Use of Digital Health Technologies in Clinical Research. *Digital Biomarkers*, 3(3), 116–132. (2019).
- ¹⁶ Baldassano, S. N., Hill, C. E., Shankar, A., Bernabei, J., Khankhanian, P., & Litt, B. Big data in status epilepticus. *Epilepsy & Behavior : E&B*, 101(Pt B), 106457. (2019).
- ¹⁷ Blair, L. M. Publicly Available Data and Pediatric Mental Health: Leveraging Big Data to Answer Big Questions for Children. *Journal of Pediatric Health Care : Official Publication of National Association of Pediatric Nurse Associates & Practitioners*, 30(1), 84–87. (2016).
- ¹⁸ Espay, A. J., Bonato, P., Nahab, F. B., Maetzler, W., Dean, J. M., Klucken, J., Eskofier, B. M., Merola, A., Horak, F., Lang, A. E., Reilmann, R., Giuffrida, J., Nieuwboer, A., Horne, M., Little, M. A., Litvan, I., Simuni, T., Dorsey, E. R., Burack, M. A., Papapetropoulos, S. Technology in Parkinson's disease: Challenges and opportunities. *Movement Disorders : Official Journal of the Movement Disorder Society*, 31(9), 1272–1282. (2016).

-
- ¹⁹ Foreman, B. Neurocritical Care: Bench to Bedside (Eds. Claude Hemphill, Michael James) Integrating and Using Big Data in Neurocritical Care. *Neurotherapeutics : The Journal of the American Society for Experimental NeuroTherapeutics*, 17(2), 593–605. (2020).
- ²⁰ Hafferty, J. D., Smith, D. J., & McIntosh, A. M. Invited Commentary on Stewart and Davis " 'Big data' in mental health research-current status and emerging possibilities". *Social Psychiatry and Psychiatric Epidemiology*, 52(2), 127–129. (2017).
- ²¹ Hemingway, H., Asselbergs, F. W., Danesh, J., Dobson, R., Maniadakis, N., Maggioni, A., van Thiel, G. J. M., Cronin, M., Brobert, G., Vardas, P., Anker, S. D., Grobbee, D. E., & Denaxas, S. (2018). Big data from electronic health records for early and late translational cardiovascular research: Challenges and potential. *European Heart Journal*, 39(16), 1481–1495. <https://doi.org/10.1093/eurheartj/ehx487>
- ²² Rodriguez, A., Smielewski, P., Rosenthal, E., & Moberg, D. Medical Device Connectivity Challenges Outline the Technical Requirements and Standards For Promoting Big Data Research and Personalized Medicine in Neurocritical Care. *Military Medicine*, 183(suppl_1), 99–104. (2018).
- ²³ Rumsfeld, J. S., Joynt, K. E., & Maddox, T. M. Big data analytics to improve cardiovascular care: Promise and challenges. *Nature Reviews. Cardiology*, 13(6), 350–359. (2016).
- ²⁴ Schofield, P. Big data in mental health research - do the ns justify the means? Using large data-sets of electronic health records for mental health research. *BJPsych Bulletin*, 41(3), 129–132. (2017).
- ²⁵ Shen, B., Lin, Y., Bi, C., Zhou, S., Bai, Z., Zheng, G., & Zhou, J. Translational Informatics for Parkinson's Disease: From Big Biomedical Data to Small Actionable Alterations. *Genomics, Proteomics & Bioinformatics*, 17(4), 415–429. (2019).
- ²⁶ Silverio, A., Cavallo, P., Rosa, R. de, & Galasso, G. Big Health Data and Cardiovascular Diseases: A Challenge for Research, an Opportunity for Clinical Care. *Frontiers in Medicine*, 6, 36. (2019).
- ²⁷ Stephenson, D., Alexander, R., Aggarwal, V., Badawy, R., Bain, L., Bhatnagar, R., Bloem, B. R., Boroojerdi, B., Burton, J., Cedarbaum, J. M., Cosman, J., Dexter, D. T., Dockendorf, M., Dorsey, E. R., Dowling, A. V., Evers, L. J. W., Fisher, K., Frasier, M., Garcia-Gancedo, L., . . . Zach, N. Precompetitive Consensus Building to Facilitate the Use of Digital Health Technologies to Support Parkinson Disease Drug Development through Regulatory Science. *Digital Biomarkers*, 4(Suppl 1), 28–49. (2020).
- ²⁸ Termine, A., Fabrizio, C., Strafella, C., Caputo, V., Petrosini, L., Caltagirone, C., Giardina, E., & Cascella, R. Multi-Layer Picture of Neurodegenerative Diseases: Lessons from the Use of Big Data through Artificial Intelligence. *Journal of Personalized Medicine*, 11(4). (2021).
- ²⁹ van den Heuvel, L., Dorsey, R. R., Prainsack, B., Post, B., Stiggelbout, A. M., Meinders, M. J., & Bloem, B. R. Quadruple Decision Making for Parkinson's Disease Patients: Combining Expert Opinion, Patient Preferences, Scientific Evidence, and Big Data Approaches to Reach Precision Medicine. *Journal of Parkinson's Disease*, 10(1), 223–231. (2020).
- ³⁰ Andy, A. U., Guntuku, S. C., Adusumalli, S., Asch, D. A., Groeneveld, P. W., Ungar, L. H., & Merchant, R. M. Predicting Cardiovascular Risk Using Social Media Data: Performance Evaluation of Machine-Learning Models. *JMIR Cardio*, 5(1), e24473. (2021).
- ³¹ Clark, R. A., Foote, J., Versace, V. L., Brown, A., Daniel, M., Coffee, N. T., Marin, T. S., Kourbelis, C., Arstall, M., Ganesan, A., Maddison, R., Kelly, J., Barry, T., Keech, W., & Nicholls, S. J. The Keeping on Track Study: Exploring the Activity Levels and Utilization of Healthcare Services of Acute Coronary Syndrome (ACS) Patients in the First 30-Days after Discharge from Hospital. *Medical Sciences (Basel, Switzerland)*, 7(4). (2019).
- ³² Haines-Delmont, A., Chahal, G., Bruen, A. J., Wall, A., Khan, C. T., Sadashiv, R., & Fearnley, D. Testing Suicide Risk Prediction Algorithms Using Phone Measurements With Patients in Acute Mental Health Settings: Feasibility Study. *JMIR MHealth and UHealth*, 8(6), e15901. (2020).

-
- ³³ Jacobson, N. C., Summers, B., & Wilhelm, S. Digital Biomarkers of Social Anxiety Severity: Digital Phenotyping Using Passive Smartphone Sensors. *Journal of Medical Internet Research*, 22(5). (2020).
- ³⁴ Li, B., Ding, S., Song, G., Li, J., & Zhang, Q. Computer-Aided Diagnosis and Clinical Trials of Cardiovascular Diseases Based on Artificial Intelligence Technologies for Risk-Early Warning Model. *Journal of Medical Systems*, 43(7), 228. (2019).
- ³⁵ Papadopoulos, A., Iakovakis, D., Klingelhoefer, L., Bostantjopoulou, S., Chaudhuri, K. R., Kyritsis, K., Hadjidimitriou, S., Charisis, V., Hadjileontiadis, L. J., & Delopoulos, A. Unobtrusive detection of Parkinson's disease from multi-modal and in-the-wild sensor data using deep learning techniques. *Scientific Reports*, 10(1), 21370. (2020).
- ³⁶ Payrovnaziri, S. N., Barrett, L. A., Bis, D., Bian, J., & He, Z. Enhancing Prediction Models for One-Year Mortality in Patients with Acute Myocardial Infarction and Post Myocardial Infarction Syndrome. *Studies in Health Technology and Informatics*, 264, 273–277. (2019).
- ³⁷ Ross, E. G., Jung, K., Dudley, J. T., Li, L., Leeper, N. J., & Shah, N. H. Predicting Future Cardiovascular Events in Patients With Peripheral Artery Disease Using Electronic Health Record Data. *Circulation. Cardiovascular Quality and Outcomes*, 12(3), e004741. (2019).
- ³⁸ Sajal, M. S. R., Ehsan, M. T., Vaidyanathan, R., Wang, S., Aziz, T., & Mamun, K. A. A. Telemonitoring Parkinson's disease using machine learning by combining tremor and voice analysis. *Brain Informatics*, 7(1), 12. (2020).
- ³⁹ Sükei, E., Norbury, A., Perez-Rodriguez, M. M., Olmos, P. M., & Artés, A. Predicting Emotional States Using Behavioral Markers Derived From Passively Sensed Data: Data-Driven Machine Learning Approach. *JMIR MHealth and UHealth*, 9(3), e24465. (2021).
- ⁴⁰ Ahn, I., Na, W., Kwon, O., Yang, D. H., Park, G. M., Gwon, H., Kang, H. J., Jeong, Y. U., Yoo, J., Kim, Y., Jun, T. J., & Kim, Y. H. Cardionet: A manually curated database for artificial intelligence-based research on cardiovascular diseases. *BMC Medical Informatics and Decision Making*, 21(1), 29. (2021).
- ⁴¹ Matoba, T., Kohro, T., Fujita, H., Nakayama, M., Kiyosue, A., Miyamoto, Y., Nishimura, K., Hashimoto, H., Antoku, Y., Nakashima, N., Ohe, K., Ogawa, H., Tsutsui, H., & Nagai, R. Architecture of the Japan Ischemic Heart Disease Multimodal Prospective Data Acquisition for Precision Treatment (J-IMPACT) System. *International Heart Journal*, 60(2), 264–270. (2019).
- ⁴² Perera, G., Broadbent, M., Callard, F., Chang, C. K., Downs, J., Dutta, R., Fernandes, A., Hayes, R. D., Henderson, M., Jackson, R., Jewell, A., Kadra, G., Little, R., Pritchard, M., Shetty, H., Tulloch, A., & Stewart, R. Cohort profile of the South London and Maudsley NHS Foundation Trust Biomedical Research Centre (SLaM BRC) Case Register: Current status and recent enhancement of an Electronic Mental Health Record-derived data resource. *BMJ Open*, 6(3), e008721. (2016).
- ⁴³ Fair Principles. [Cited 7 July 2022]. Available from: <https://www.go-fair.org/fair-principles/>
- ⁴⁴ Data Access Quality and Curation for Observational Research Design. [Cited 7 July 2022]. Available from: <https://www.daqcord.org/>
- ⁴⁵ Cerreta F, Ritzhaupt A, Metcalfe T, Askin S, Duarte J, Berntgen M & Vamvakas S. Digital technologies for medicines: shaping a framework for success. *Nat Rev Drug Discov* (9):573-574. (202).
- ⁴⁶ FHIR. [Cited 7 July 2022]. Available from: <https://www.hl7.org/fhir/index.html>
- ⁴⁷ SNOMED. [Cited 7 July 2022]. Available from: <https://www.snomed.org/>
- ⁴⁸ Shi, P.; Cui, Y.; Xu, K.; Zhang, M. & Ding, L. Data Consistency Theory and Case Study for Scientific Big Data. *Information* 10, 137. (2019).
- ⁴⁹ Delgado-Rodríguez, M. & Llorca, J. Bias. *Journal of Epidemiology & Community Health* 58. 635-641. (2004).
- ⁵⁰ Freudenheim, J.L., Ritz, J., Smith-Warner, S.A., Albanes, D., Bandera, E.V., van den Brandt, P.A., Colditz, G., Feskanich, D., Goldbohm, R.A., Harnack, L., Miller, A.B., Rimm, E., Rohan, T.E., Sellers, T.A., Virtamo, J., Willett, W. C. & Hunter, D.J. Alcohol consumption and risk of lung cancer:

-
- a pooled analysis of cohort studies. *The American Journal of Clinical Nutrition* 82, 3, 657–667. (2005).
- ⁵¹ Altman, D. G., Vergouwe, Y., Royston, P. & Moons, K. G. M. Prognosis and prognostic research: validating a prognostic model. *BMJ* 338. (2009).
- ⁵² Ranstam, J. Methodological Note: Accuracy, precision, and validity. *Acta Radiologica* 49:1. 105-106. (2008).
- ⁵³ Trajković, G. Measurement: Accuracy and Precision, Reliability and Validity. In: Kirch W. (eds) *Encyclopedia of Public Health*. Springer, Dordrecht. (2008).
- ⁵⁴ National Academies of Sciences, Engineering, and Medicine. *Reproducibility and Replicability in Science*. Washington, DC: The National Academies Press. Chap. 3. (2019)
- ⁵⁵ Kukull, W.A. & Ganguli, M. Generalizability: the trees, the forest, and the low-hanging fruit. *Neurology* 78(23):1886-91. (2012).
- ⁵⁶ D'Souza, R.S., Hooten, W.M. & Murad, M.H. A Proposed Approach for Conducting Studies That Use Data From Social Media Platforms. *Mayo Clin Proc* 96(8):2218-2229. (2021).
- ⁵⁷ Kim, Y., Huang, J. & Emery, S. Garbage in, Garbage Out: Data Collection, Quality Assessment and Reporting Standards for Social Media Data Use in Health Research, Infodemiology and Digital Disease Detection. *J Med Internet Res* 18(2):e41. (2016).
- ⁵⁸ Open industry specifications, models and software for e-health. [Cited 7 July 2022]. Available from: <https://www.openehr.org/>
- ⁵⁹ Jones, K.H., Ford, E.M., Lea, N., Griffiths, L.J., Hassan, L., Heys, S., Squires, E. & Nenadic, G. Toward the Development of Data Governance Standards for Using Clinical Free-Text Data in Health Research: Position Paper. *J Med Internet Res* 22(6):e16760. (2020).
- ⁶⁰ Digital Medicine Society. [Cited 7 July 2022]. Available from: <https://www.dimesociety.org/>
- ⁶¹ Bradway, M., Gabarron, E., Johansen, M., Zanaboni, P., Jardim, P., Joakimsen, R., Pape-Haugaard, L. & Årsand, E. Methods and Measures Used to Evaluate Patient-Operated Mobile Health Interventions: Scoping Literature Review, *JMIR Mhealth Uhealth* 8(4):e16814. (2020).
- van de Leur, R.R., Boonstra, M.J., Bagheri, A., Roudijk, R.W., Sammani, A., Taha, K., Doevendans, P.A., van der Harst, P., van Dam, P.M., Hassink, R.J., van Es, R. & Asselbergs, F.W. Big Data and Artificial Intelligence: Opportunities and Threats in Electrophysiology. *Arrhythm Electrophysiol Rev.* 9(3):146-154. (2020).
- ⁶² Luo, W., Phung, D., Tran, T., et al. Guidelines for Developing and Reporting Machine Learning Predictive Models in Biomedical Research: A Multidisciplinary View. *J Med Internet Res* 18(12):e323. (2016)
- ⁶³ STROBE. Strengthening the reporting of observational studies in epidemiology. [Cited 7 July 2022]. Available from: <https://www.strobe-statement.org/>