Knowledge graphs, clinical trials, dataspace, and AI: Uniting for progressive healthcare innovation

Knowledge Graphs, Clinical Trials, Dataspace, and AI: Uniting for Progressive Healthcare Innovation

Mohan Timilsina
Insight Centre for Data Analytics
University of Galway
Galway, Ireland
mohan.timilsina@insight-centre.org

Saeed Alsamhi
Insight Centre for Data Analytics
University of Galway
Galway, Ireland
saeed.alsamhi@insight-centre.org

Rafiqul Haque
Insight Centre for Data Analytics
University of Galway
Galway, Ireland
rafiqul.haque@insight-centre.org

Conor Judge
College of Medicine, Nursing, Health Sciences
University of Galway
Galway, Ireland
conor.judge@universityofgalway.ie

Edward Curry
Insight Centre for Data Analytics
University of Galway
Galway, Ireland
edward.curry@insight-centre.org

***Abstract*—Amidst prevailing healthcare challenges, a dynamic solution emerges, fusing knowledge graph technology, clinical trials optimization, dataspace integration, and AI innovation. This unified approach tackles issues like limited patient insights, suboptimal trial designs, and imprecise treatments. By interlinking diverse data through knowledge graphs, this method illuminates disease trends, therapeutic efficacies, and patient prognoses. AI techniques, especially machine learning, contribute predictive power by unveiling hidden patterns for accurate diagnostics, prognostics, and personalized treatments. This multidisciplinary fusion transforms clinical trials, enhancing comprehensiveness and precision through real-world data analysis and subgroup identification. In reshaping healthcare, this proposition aims to accelerate treatment personalization, elevate therapeutic efficacy, and empower informed medical decisions, encompassing the essence of 'Advancing Healthcare through Innovation: Knowledge Graphs, Clinical Trials, Dataspace, and AI'.**

***Index Terms*—dataspace, linked, clinical, machine learning, knowledge graph**

I. Introduction

The term "dataspace" can have different meanings depending on the context. In the context of the European Health Data Space (EHDS), it refers to a proposed infrastructure for the primary and secondary use of electronic health data across borders [1]. The need for a dataspace can vary depending on the context. In the case of the European Health Data Space (EHDS), it is proposed to facilitate the sharing of electronic health data across borders while ensuring data quality and privacy. The EHDS aims to establish common standards and practices, infrastructures, and a governance framework for the primary and secondary use of electronic health data. The proposed infrastructure is intended to enable researchers and healthcare professionals to access and use health data from different countries to improve medical research and patient care.

Similarly, the establishment of a medical dataspace is crucial in today's healthcare landscape due to several key reasons. Firstly, healthcare generates vast amounts of data from various sources such as electronic health records, medical imaging, genomic sequencing, wearable devices, and patient-generated data. These diverse datasets hold valuable information that, when integrated and analyzed, can provide comprehensive insights into disease patterns, treatment effectiveness, and patient outcomes. Secondly, the integration and interoperability of healthcare data are essential for facilitating collaborative research and innovation. A medical dataspace enables researchers, clinicians, and data scientists to access and share data seamlessly, promoting cross-disciplinary collaborations and accelerating medical advancements [2]. Additionally, a dataspace approach can enhance the efficiency and effectiveness of clinical trials by enabling the analysis of larger datasets from real-world settings [3], leading to more comprehensive observations of treatment outcomes and identification of patient subgroups that may respond differently to interventions. Ultimately, a medical dataspace holds the potential to transform healthcare delivery, empower evidence-based decision-making, and improve patient care outcomes [4]. The harmony between clinical trial data, dataspace and artifical intelligence (AI) / machine learning(ML)[1] is depicted in the picture below:

In Figure 1, we have three main components: Clinical Trial Datasets, Dataspace and Machine Learning.

- **Clinical trial data:** It is the raw data that is collected from the clinical trials. It can inculde information about patient's demographics, medical history, treatment and outcomes.
- **Graph Dataspace:** The graph dataspace serves as a central repository for various healthcare data, including electronic health records, genomic data, wearable devices, and other health-related information. It facilitates data integration and preprocessing, ensuring data consistency

[1]Note that the words "artificial intelligence (AI)" and "machine learning (ML)" are interchangeably used in the paper.

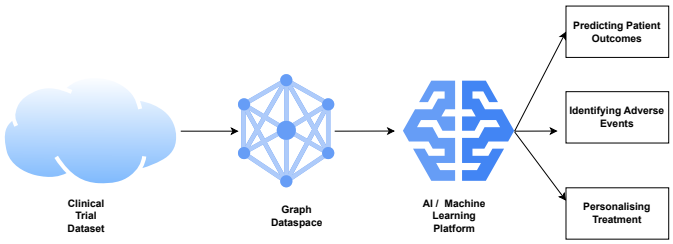

Fig. 1. The diagram illustrates the harmonious connection between clinical trial data, graph dataspace, and machine learning. The transparent blue cloud represents the diversity of data collected during clinical trials. The graph signifies the openness and accessibility of the integrated data. The hexagon structure denotes the power and sophistication of machine learning, enabling three crucial applications in clinical trials: predicting patient outcomes, identifying adverse events, and personalizing treatment.

and interoperability. This will make it more accessible to researchers and clinicians.

- **AI/Machine Learning (ML) Platform:** Machine Learning algorithms are applied to the data within the Dataspace. This step involves data analysis and insights generation, where patterns, relationships, and trends are identified. It brings together the power of the vast interconnected data sources within the graoh dataspace to drive advancements in healthcare and medical research. The insights obtained from the ML analysis can inform the design and optimization of future clinical trials. Additionally, the real-world data generated during clinical trials can be used to validate and refine the ML models, leading to a feedback loop that improves the overall healthcare innovation process.

Outline: The rest of the paper is organized in a rather standard way: motivation, problem statement, proposed solution, case study, related work, future directions and conclusion.

II. Motivation

Creating and managing a linked graph dataspace in the context of clinical trials presents several significant challenges. Firstly, data heterogeneity poses a hurdle, as clinical trial data originates from diverse sources like electronic health records, wearable devices, and genomic data. Integrating these distinct data types into a coherent graph structure requires addressing differences in formats, semantics, and quality. Ensuring data quality and accuracy is another formidable challenge. Clinical trial data can be noisy, incomplete, or inconsistent, potentially leading to erroneous conclusions. Maintaining data integrity across various nodes and relationships is crucial to obtain reliable insights. Privacy and security concerns also loom large. Clinical trial data often contains sensitive patient information, necessitating strict compliance with privacy regulations like HIPAA[2] or GDPR[3]. Balancing data utility with stringent privacy measures poses a continuous challenge. The dynamic nature of clinical data further complicates matters. Clinical trial data is subject to updates, additions, and revisions over time. Ensuring the graph remains up-to-date while preserving historical context requires robust version control mechanisms. Interoperability is yet another challenge. Different clinical trials employ varying data models, making it difficult to seamlessly integrate and compare data across trials. A linked graph dataspace must bridge these data model disparities to facilitate cross-trial analysis effectively. Lastly, scalability and computational efficiency are essential. As clinical trial datasets grow larger, managing and analyzing the linked graph becomes more resource-intensive. Efficient algorithms and infrastructure are necessary to support timely queries and analyses without compromising performance. Addressing these challenges requires multidisciplinary collaboration, incorporating expertise from data science, domain-specific knowledge, and regulatory compliance to ensure the successful establishment and utilization of a linked graph dataspace for clinical trials.

[2]https://www.hhs.gov/hipaa/index.html
[3]https://gdpr-info.eu/

III. Problem Statement

In the field of healthcare and clinical research, the management and analysis of data have become increasingly intricate due to the sheer volume, diversity, and complexity of available information. Traditional clinical dataspace systems often struggle to effectively integrate and derive meaningful insights from these diverse data sources, including electronic health records, genomic data, treatment records, and patient demographics. This limitation hampers the potential for comprehensive data-driven decision-making, hindering the progress of personalized medicine, predictive modeling, and innovative treatment strategies.

A. Proposed Solution

The conventional clinical dataspace often struggles to holistically capture the intricate interconnections within diverse healthcare data, limiting the potential for comprehensive insights and informed decision-making. The linked graph dataspace emerges as a solution by harnessing graph-based representation, enabling the portrayal of complex relationships between patients, diseases, treatments, biomarkers, and more. This approach not only enhances data integration and quality but also empowers researchers to uncover nuanced correlations, fostering personalized medicine, predictive analytics, and optimized treatment strategies for improved patient outcomes. The overall idea is demonstrate in the Figure 2.

B. Graph Processing

Graph processing via RDF (Resource Description Framework) in healthcare dataspace offers a powerful approach to representing and analyzing complex healthcare data [5]. RDF is a standardized data model used to describe resources and their relationships in a graph format, making it an ideal technology for modeling diverse healthcare data sources and their interconnections [6]. In the context of healthcare dataspace, RDF facilitates the integration of various types of data, including electronic health records (EHRs), patient demographics, genomic data, treatment information, and more, into a unified knowledge graph. The key advantage of using

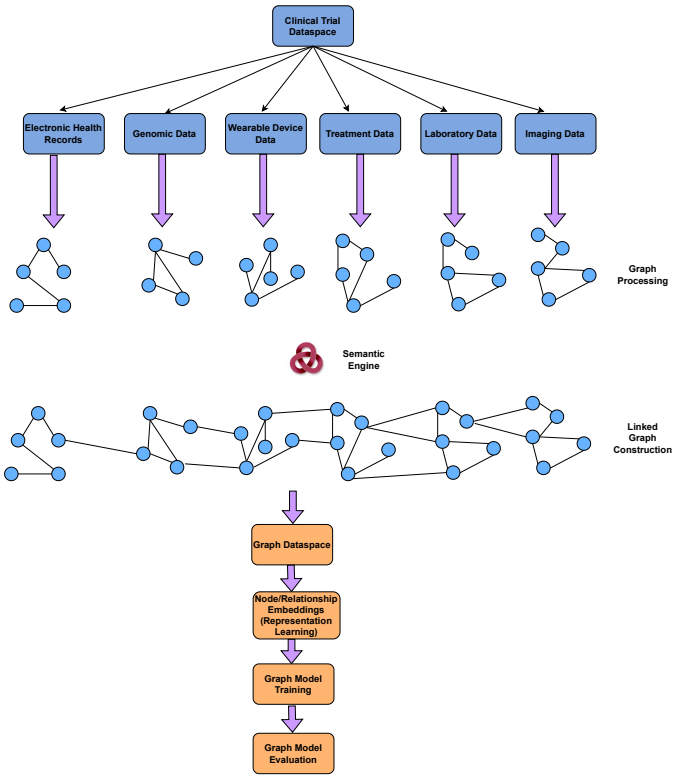

Fig. 2. Solution Approach.

RDF in healthcare dataspace is its ability to capture and represent semantic relationships between different entities. Each piece of data, or resource, is represented as a node in the graph, and the relationships between these nodes are expressed through edges. This allows for the creation of a rich and interconnected network of data that goes beyond traditional tabular or relational data models. RDF's flexibility in capturing complex relationships enables the integration of data from different sources, making it a valuable tool for creating a comprehensive view of a patient's health profile [7]. Graph processing via RDF also enables sophisticated query capabilities and data analytics. SPARQL (SPARQL Protocol and RDF Query Language) is the query language used with RDF, which allows for the retrieval and analysis of specific patterns within the graph [8]. Researchers and healthcare professionals can use SPARQL to ask complex questions and gain valuable insights from the integrated data. For example, they can perform patient cohort analyses, identify patterns of disease occurrence, study the effectiveness of treatments, and predict patient outcomes based on the collective knowledge present in the healthcare dataspace. Furthermore, RDF-based healthcare dataspace can be leveraged for decision support systems, clinical trials, and precision medicine initiatives [9]. RDF's ability to handle heterogeneous and multi-modal data ensures that diverse types of information can be utilized together to derive meaningful and personalized insights. This is particularly important in precision medicine, where a patient's genetic, lifestyle, and clinical data can be combined to create tailored treatment plans.

C. Semantic Engine

A semantic engine is a specialized software system that uses semantic technologies to process, organize, and understand data in a way that adds meaning and context to the information [10]. It leverages semantic web standards, such as RDF (Resource Description Framework) and ontologies, to represent data in a structured and interconnected manner, enabling machines to comprehend and reason about the relationships between different pieces of information [11]. The core function of a semantic engine is to annotate data with semantic metadata, which includes standardized terms and concepts from domain-specific ontologies. This process enriches the data by providing explicit information about its meaning and context, allowing the engine to infer and deduce knowledge from the relationships between data elements [12]. Semantic engines in healthcare can help integrate electronic health records, genomic data, and medical literature into a unified knowledge graph, enabling better decision support for clinicians and researchers. Therefore semantic engine makes information more accessible, meaningful, and actionable for both humans and machines. It plays a crucial role in enhancing data understanding, interoperability, and knowledge representation in the age of big data and interconnected health information systems.

D. Linked Graph Construction

Linked Graph Construction using graph processing and semantic engines involves leveraging advanced technologies to create a connected and semantic-rich network of clinical trial data elements. Graph processing techniques, such as RDF and SPARQL (SPARQL Protocol and RDF Query Language), enable efficient traversal and analysis of relationships in the graph. Semantic engines play a crucial role in annotating the data with meaningful metadata, such as domain-specific ontologies and vocabularies, ensuring that each data element's context and semantics are explicitly captured. As a result, the linked graph construction creates a comprehensive knowledge graph that interconnects diverse clinical trial data sources, including electronic health records, genomic data, patient demographics, and treatment information. This interconnected knowledge graph facilitates sophisticated data querying, reasoning, and knowledge discovery, empowering researchers and healthcare professionals to gain valuable insights, optimize treatment approaches, and accelerate medical discoveries. By combining graph processing and semantic engines, linked graph construction enhances the integration and interoperability of clinical trial data, leading to more precise and personalized healthcare practices.

E. Graph Dataspace

Graph dataspace refers to a dynamic and interconnected ecosystem that encompasses a wide array of data sources and their relationships represented as graphs. In this context, the graph serves as a powerful visualization and analytical

tool, depicting entities as nodes and their associations as edges. Graph dataspace can encompass diverse data types, including structured and unstructured data, offering a holistic perspective on intricate relationships and dependencies that traditional tabular data representation might miss. Graph dataspace finds applications across various domains, including healthcare, where it proves especially valuable due to the complex interplay of medical information. In the healthcare sector, a graph dataspace could integrate electronic health records, patient demographics, medical procedures, diagnoses, medications, and more, all interconnected based on real-world relationships. This interconnectedness allows for a deeper understanding of patient profiles, disease progression, treatment effectiveness, and adverse events. The utilization of graph dataspace extends beyond visualization. It enables sophisticated querying, advanced analytics, and predictive modeling. Graph-based machine learning techniques can be employed to uncover hidden patterns, make accurate predictions, and recommend tailored treatment options. Furthermore, graph dataspace fosters collaboration, as it enables diverse stakeholders, from researchers to clinicians, to access and interpret complex data in an intuitive and meaningful way.

F. Node/Relationship Embeddings(Representation Learning)

Node and relationship embedding representation in the context of clinical graph dataspace is a sophisticated approach that enhances the understanding and utilization of complex medical data. In this methodology, nodes and their relationships within the graph are transformed into continuous, numerical vectors through embedding techniques. These embeddings capture the latent features and contextual information of nodes and edges, enabling the translation of complex relationships into a format understandable by machine learning algorithms. In clinical graph dataspace, this embedding representation holds tremendous potential. Nodes representing various medical entities, such as patients, diseases, medications, and procedures, can be transformed into dense vectors that encapsulate their inherent characteristics. Edges connecting these nodes encode the relationships, such as patient-doctor interactions, treatment associations, and disease correlations. These embeddings preserve intricate patterns that traditional tabular data formats might overlook, leading to a deeper understanding of patient profiles, disease networks, and treatment pathways. Node and relationship embedding representation also facilitates downstream machine learning tasks. By embedding nodes into continuous vectors, these representations serve as rich inputs for various algorithms like Graph Convolutional Networks (GCNs) [13], Graph Neural Networks (GNNs) [14] and and Graph Attention Networks (GATs) [15]. These algorithms can then uncover hidden patterns, predict patient outcomes, identify optimal treatment strategies, and even recommend personalized interventions based on the learned embeddings. The advantages of node and relationship embedding representation are clear: it bridges the gap between complex graph structures and machine learning algorithms, enabling the extraction of meaningful insights from clinical dataspace. By translating intricate relationships into actionable numerical representations, it empowers healthcare professionals, researchers, and data scientists to make informed decisions, drive innovation, and ultimately enhance patient care within the dynamic landscape of healthcare data.

G. Graph Model Training

Graph model training is a vital stage in the domain of graph-based machine learning, focusing on extracting valuable insights and predictive capabilities from intricate data relationships. Employing sophisticated algorithms like Graph Convolutional Networks (GCNs) and Graph Neural Networks (GNNs, this process involves iteratively refining node and edge embeddings to capture intricate patterns within the interconnected data. In healthcare, this approach gains significance as it enables the discovery of hidden correlations, predictive trends, and personalized patient profiles within complex medical datasets. By integrating both the graph's structure and attribute information, these models excel in disease classification, outcome prediction, and treatment optimization. Continuously learning from evolving data, graph model training holds transformative potential, offering healthcare professionals enhanced diagnostic tools, early detection algorithms, and personalized intervention strategies that adapt to the dynamic nature of medical knowledge.

H. Graph Model Evaluation

Graph model evaluation is a crucial step in the realm of graph-based machine learning, ensuring the accuracy, robustness, and generalizability of predictive models constructed from complex data relationships. After training graph models using algorithms, their performance is rigorously assessed. This process involves testing the models on unseen data and assessing their ability to make accurate predictions, classify diseases, or recommend treatments based on the graph's interconnected structure. In the context of healthcare, graph model evaluation is of paramount importance. It enables healthcare professionals to gauge the model's ability to identify intricate correlations, anticipate patient outcomes, and provide relevant insights. A well-evaluated graph model ensures reliable decision support and contributes to evidence-based medical practices. By benchmarking these models against established metrics, such as precision, recall, AUC-ROC and F1-score, their real-world applicability and contribution to healthcare advancement can be comprehensively measured. As graph models continue to evolve and adapt to changing data, effective evaluation ensures that these models remain valuable tools for enhancing diagnostic accuracy, predicting patient outcomes, and driving informed clinical decisions.

IV. Case Study

A. Graph-Based Medical Dataspace for Predicting Relapse in Early-Stage NSCLC Patients

Introduction: In the field of oncology, personalized treatment strategies have gained prominence in improving patient outcomes. One key aspect is predicting the risk of relapse for

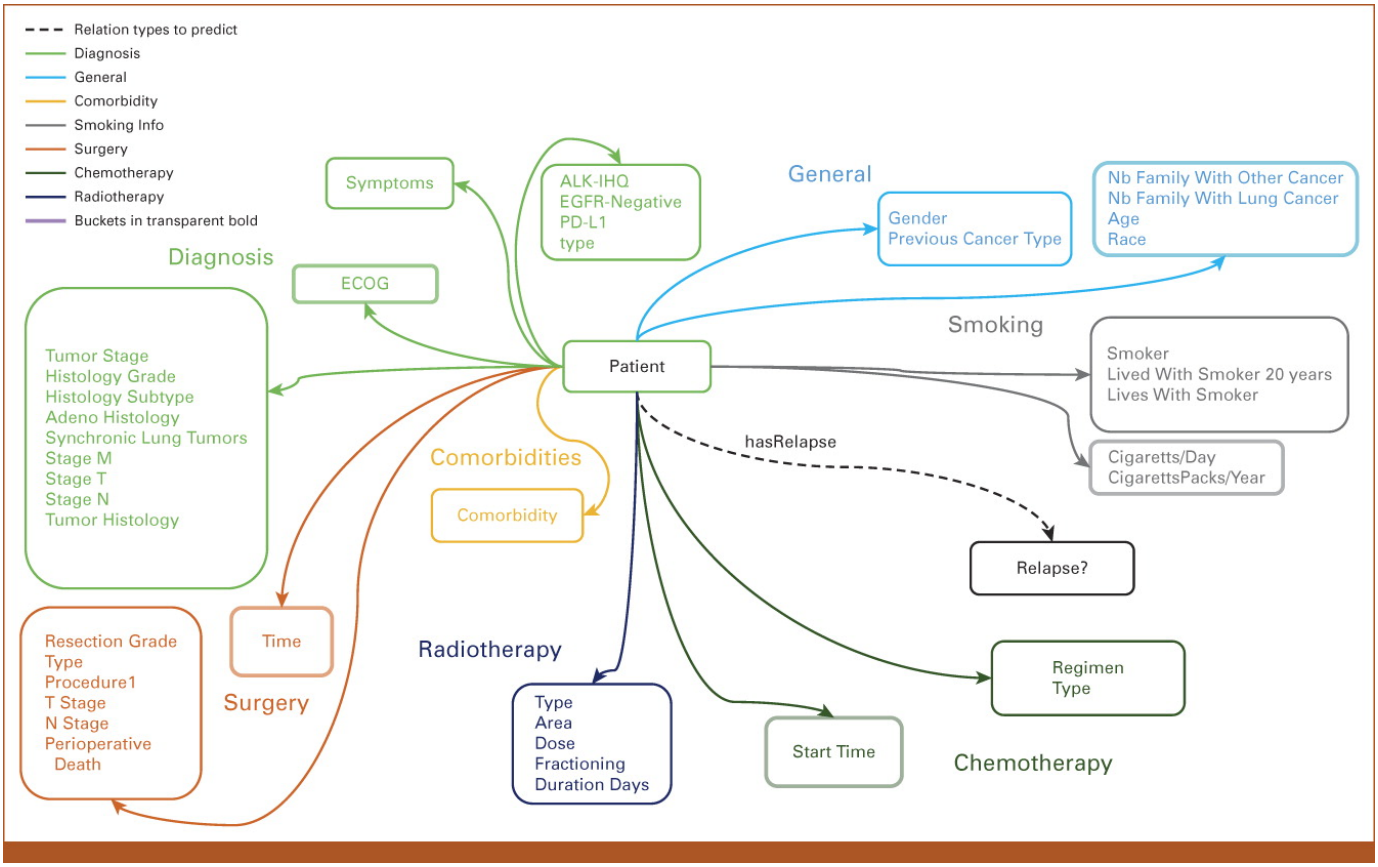

Fig. 3. Schema of the clinical data modeled as a knowledge graph [16].

patients with early-stage non–small-cell lung cancer (NSCLC). Machine learning techniques offer a promising avenue for accurate prediction and personalized care. In this case study, we explore the use of a graph-based medical dataspace to predict the probability of relapse in early-stage NSCLC patients.

Graph Dataspace Creation: We construct a graph-based dataspace that represents patient data, features, and relapse predictions. Each patient is represented as a node, linked to various features such as age, gender, and relevant medical information. Additionally, each patient node is connected to a prediction node, which holds the predicted probability of relapse. The graph's structure enables us to capture the intricate relationships between patient characteristics and relapse outcomes. Figure 3, demonstrates the knowledge graph created with consulting the clinician [16] for this study.

The creation of the graph involves defining entities, their specific attributes, and the relationships that interconnect them. In the context of this study, these entities encompass patients, diseases, and biomarkers, with their interactions visualized as links within the graph structure. For example, a patient entity can possess attributes such as age, gender, and smoking history, while being linked to corresponding disease entities if applicable. Diverse and varied data sources are harnessed, constituting clinical information like patient demographics (age, sex, TNM stage), treatment specifics (surgery type, time elapsed since diagnosis), and genetic insights (e.g., mutations like EGFR and KRAS relevant to lung cancer). The knowledge graph serves as an integrative platform, harmonizing these disparate data elements into a coherent and interconnected whole. This comprehensive knowledge graph encompasses a staggering 34,351 connections, intricately linking nodes across the entire dataset. It comprehensively encompasses the clinical records of 1,387 patients diagnosed with early-stage NSCLC. This meticulously constructed graph then becomes the foundation for training and evaluating machine learning models specifically designed for forecasting the likelihood of relapse in NSCLC patients.

Results: The graph-based machine learning algorithm called ComplEx-N3 (achieving state-of-the art results [17]), achieved a 68% accuracy in predicting the probability of relapse in patients with early-stage non-small cell lung cancer (NSCLC) over a held-out test set of 200 patients, calibrated on a held-out set of 100 patients. The model leverages dependencies and patterns between patients and other concepts in the graph, such as diseases, biomarkers, genes, and drugs, to deliver more powerful predictions over traditional approaches. The graph-based model is well-suited to be used by various reasoners, including logical, machine learning-based, or hybrid, and allows inference over long-range dependencies between concepts.

Overall, the results suggest that the graph-based machine

learning model can enable objective, personalized, and reproducible prediction of relapse and, therefore, disease outcome in patients with early-stage NSCLC. However, the study is limited to a cohort of patients diagnosed and treated in Spain, and further prospective and multisite validation is needed to assess the generalizability of the model.

V. Related Work

In this section, we review the work that tackles the same problems as this paper, the methods used, and their strengths and weaknesses. In Figure 4, we have identified potential topics that contribute towards harnessing Dataspace, Machine Learning, and Clinical Trials. Below, we discuss each of these topics.

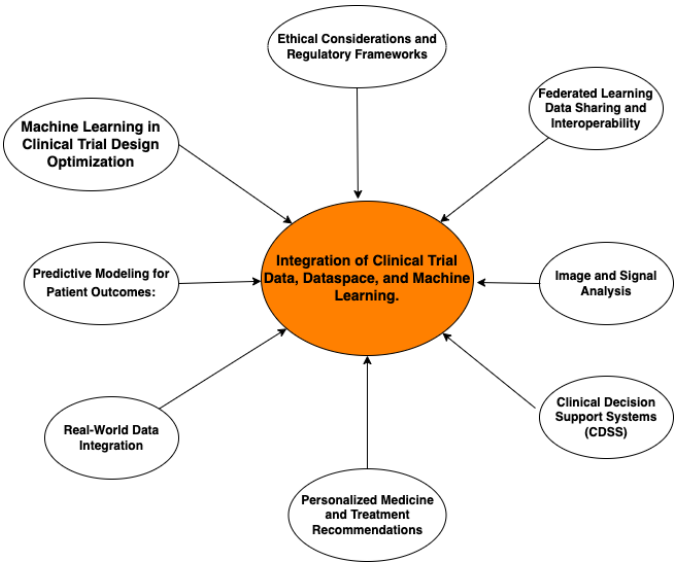

Fig. 4. The diagram highlights key research areas in integrating clinical trial data, dataspace, and machine learning in healthcare.

A. Machine Learning in Clinical Trial Design Optimization

Studies have explored the application of machine learning algorithms to optimize various aspects of clinical trial design, such as patient recruitment, stratification, and endpoint selection. This optimization may involve leveraging data from the dataspace to inform trial design decisions. Clinical research encompasses both interventional and observational studies that involve volunteers to assess the efficiency and safety of new drugs, treatments, or medical devices [18]. An optimized trial design process is crucial for ensuring a successful clinical study, efficient utilization of resources, and ethical treatment of participants [19]. The growing interest in AI (Artificial Intelligence) and ML in healthcare has primarily centered around healthcare delivery, with less emphasis on clinical research [20]. Existing studies have largely concentrated on predicting trial outcomes rather than optimizing the design of clinical trials [21], [22]. However, by harnessing the potential of historical data through machine learning, there is an opportunity to streamline the trial design process, leading to reduced time and cost expenditures. Kavalci et al [23] generated an enhanced clinical trial dataset by transforming and integrating two public datasets of from various diseases and conditions, and proposed using a set of new features regarding eligibility criteria and disease categories. The authors used machine learning models to predict early termination for any new interventional trial protocol using the enhanced clinical trial dataset generated in this study. In terms of ML, the limitation of this study is that the dataset used to generate eligibility criteria search features may not be comprehensive enough to cover all the eligibility criteria categories and entities for earlier phase studies which might impact the performance of the model. By analyzing data from previous trials, ML can help researchers optimize the choice of treatment regimens for testing and identify potential problems that could prevent a trial from being successful. While these approaches show promise, more research is needed to determine how effective they are.

B. Predictive Modeling for Patient Outcomes

Machine learning models have been developed to predict patient outcomes based on various clinical and demographic factors. These models may use data from the dataspace to improve their predictive accuracy and generalizability. Predictive modeling is a powerful tool that can be used to improve patient outcomes. By using historical data, predictive models can be built to identify patients who are at risk of developing certain diseases or experiencing adverse drug reactions [24]–[26]. This information can then be used to intervene early and prevent negative outcomes. The digitization of healthcare data has enabled the large-scale analysis of healthcare data for public health activities such as pharmacovigilance. Electronic health records of the patients are a valuable data source for pharmacovigilance, but there are challenges in representing the heterogeneous data types in a way that is conducive to machine learning [27]. In such situation, dataspaces can offer help. For instance, dataspaces are a way of organizing and managing heterogeneous data from different sources. They provide a common framework for accessing and sharing data, and they can help to overcome the challenges of representing high-dimensional and sparse data. Similarly, for predictive modeling for patient outcomes, dataspace would allow for the integration of real-world clinical data into disease classification and care pathways, leading to appropriate patient stratification and optimal care [28]. Therefore, the combination of predictive modeling and dataspaces has the potential to revolutionize the way that healthcare is delivered. By providing healthcare professionals with more accurate and timely information, dataspaces can help to improve patient outcomes and reduce the cost of healthcare.

C. Real World Data Integration using Knowledge Graph

In the context of clinical trials, the emergence of a dynamic solution hinges on the pivotal role of knowledge graphs in orchestrating the integration of real-world data. This innovative approach stands as a potent response to the challenges

presented by the incorporation of diverse and complex data streams. By harnessing the capabilities of knowledge graph technology addresses the limitations associated with traditional trial designs, where siloed data sources often hinder a comprehensive understanding of patient dynamics and treatment outcomes [29]. Knowledge graphs act as the central link that weaves together different types of data like electronic health records, genomics, patient reports, and wearable device data. This connection creates a complete picture of patient health and treatment progress, breaking down barriers between different sets of data [30]. In the domain of clinical trials, where precision and agility are paramount, knowledge graphs enable the fusion of diverse data modalities into a cohesive narrative. This holistic representation fosters a deeper comprehension of patient responses to interventions, unearthing nuanced insights that might remain concealed in isolated data pockets. By leveraging AI techniques, particularly machine learning, knowledge graphs unravel hidden patterns, anticipate patient trajectories, and facilitate the identification of distinct patient cohorts that might respond differentially to treatments [31]. Moreover, the symbiotic relationship between knowledge graphs and clinical trials ushers in a transformative era of real-world data integration. It empowers researchers and clinicians to navigate the intricate patient landscape with unprecedented clarity, resulting in accelerated decision-making, tailored interventions, and elevated therapeutic precision. This innovation encapsulates the essence of harnessing knowledge graphs for "real-world data integration" within clinical trials, culminating in an enriched understanding of patient health, streamlined research endeavors [32].

D. Clinical Decision Support Systems (CDSS)

Clinical Decision Support Systems (CDSS) and dataspace are two interconnected components that play a pivotal role in enhancing healthcare delivery and decision-making. CDSS is a technology-driven system that assists healthcare professionals in making informed and evidence-based decisions at the point of care [33]. It leverages patient data, medical knowledge, and algorithms to provide timely and context-specific recommendations, alerts, and reminders to clinicians [34]. The synergy between CDSS and dataspace amplifies their respective benefits, resulting in improved clinical decision-making [26], [35]–[37]. By integrating CDSS with dataspace, clinicians gain access to a broader and richer set of patient data. This empowers CDSS to deliver more accurate and personalized recommendations, taking into account a patient's entire health history, genetic information, and real-time data [38]. Dataspace provides a valuable repository of evidence-based medical knowledge and guidelines, which CDSS can leverage to generate more relevant and up-to-date clinical decision support. The integration of diverse datasets within dataspace allows CDSS to draw insights from real-world patient outcomes, enabling better-informed treatment recommendations and care plans [38]. Moreover, the integration of CDSS with dataspace facilitates continuous learning and improvement. As CDSS interacts with patient data within the dataspace, it can adapt and refine its recommendations over time, aligning with evolving medical research and best practices. However, synergy between CDSS and dataspace also presents challenges, such as ensuring data privacy and security, addressing data interoperability, and maintaining the accuracy and reliability of the underlying data sources [39], [40]. By addressing these challenges and fostering collaborative efforts between healthcare stakeholders, CDSS and dataspace can revolutionize clinical decision-making, leading to more efficient, personalized, and evidence-based healthcare practices.

E. Image and Signal Analysis

Image and Signal Analysis plays a crucial role in healthcare research and clinical trials, enabling the extraction of valuable information from medical images and signals generated by various devices [41]. Integrating this analysis with "clinical trial data, dataspace, and machine learning" yields a powerful and comprehensive approach to improving patient outcomes and advancing medical research. In clinical trials, "Image and Signal Analysis" can be applied to medical images, such as X-rays, MRI scans, and histopathological slides, to identify disease markers, track disease progression, and evaluate treatment responses [42]. The data obtained from these analyses can be combined with "clinical trial data," including patient demographics, medical history, and treatment details, to create a rich and diverse dataset within "dataspace." With the integration of "machine learning" algorithms, this multimodal dataset can be harnessed to develop predictive models and decision support systems [43]. Machine learning models can identify patterns, correlations, and predictive markers that might not be apparent through manual analysis alone. By leveraging these advanced techniques, researchers can gain deeper insights into treatment efficacy, patient stratification, and potential adverse events. The "integration of clinical trial data, dataspace, and machine learning" promotes evidence-based decision-making and facilitates real-time data analysis during trials [20], [22]. Researchers can adapt trial protocols based on emerging patient data, optimizing treatment strategies for improved patient outcomes. Moreover, the integration fosters collaborative research by enabling data sharing and analysis across different trials and healthcare institutions. This collective knowledge can lead to more accurate and generalizable findings, advancing medical knowledge and enhancing the overall efficiency of clinical trials [20], [44]. However, challenges exist in this integration, including ensuring data privacy and security, managing large and diverse datasets, and developing robust machine learning models that are interpretable and reliable [45]–[47]. Overcoming these challenges requires multidisciplinary collaboration and a strong focus on data governance and standardization.

F. Federated Learning Data Sharing and Interoprability

Federated Learning [48], Data Sharing [49], and Interoperability [50] is a cutting-edge approach in healthcare that can significantly enhance the integration of clinical trial data, dataspace, and machine learning. This novel framework

addresses the challenges of data privacy and data sharing while promoting collaborative research and advancing medical knowledge. In a federated learning setting, data from multiple healthcare institutions or clinical trial sites remain decentralized, and the raw data remains within their respective premises [51]. Instead of centralizing the data in a single location, federated learning enables the sharing of machine learning models across institutions. These models are then trained locally on the distributed data, and only the model updates, rather than raw data, are exchanged between sites. This approach ensures data privacy and security while allowing for collaborative model training. By integrating "clinical trial data" with the federated learning framework, researchers can tap into a broader and more diverse dataset, encompassing multiple sites' patient populations. This integration enhances the statistical power and generalizability of machine learning models, leading to more accurate and reliable predictions and treatment recommendations [52]. Moreover, the incorporation of "dataspace" into the federated learning process promotes interoperability among disparate data sources [53]. Dataspace provides a standardized framework for data representation and aggregation, enabling seamless data integration from different clinical trial sites and healthcare institutions [54]. This interoperability allows researchers to combine structured and unstructured data, such as electronic health records, medical images, and patient-generated data, to create a comprehensive and unified dataset for analysis. The integration of clinical trial data, dataspace, and federated learning fosters a collaborative and ethical approach to research. Healthcare institutions can participate in research projects without compromising patient privacy, as sensitive data remains locally stored and not shared in its raw form [55]. This encourages data sharing and cross-institutional collaborations, accelerating medical discoveries and advancements. However, this integration also comes with challenges. Ensuring data standardization across different sites, managing data quality, and coordinating model updates can be complex tasks. Additionally, the federated learning approach may require overcoming technical hurdles to handle varying data types and distributions.

G. Ethical Considerations and Regulatory Frameworks

The integration of clinical trial data dataspace and machine learning has the potential to improve the efficiency and effectiveness of clinical trials, as well as to generate new insights into disease and treatment. However, it is important to carefully consider the ethical and regulatory implications of these technologies before they are widely adopted [20], [56]. Some of the ethical considerations that need to be taken into account include *privacy* and *confidentiality*, *fairness* and *non-discrimination*, and *transparency* and *accountability*. The General Data Protection Regulation (GDPR) [4] is a European Union regulation that sets out rules for the processing of personal data, including health data. The European Health Data Space (EHDS) [5] is a proposed framework for the sharing of health data across the European Union. The EHDS would include provisions for protecting the privacy and confidentiality of health data. Alongside EHDS, the Towards the European Health Data Space (TEHDAS) [6] initiative further accelerates the development of a federated and collaborative health data ecosystem. TEHDAS promotes the exchange of expertise and best practices across EU member states, facilitating seamless data sharing while preserving data privacy and security. In addition to these ethical considerations, there are a number of other factors that need to be considered when integrating clinical trial data dataspace and machine learning. These factors include the accuracy of the data, the robustness of the machine learning models, and the potential impact of the use of these technologies on individuals and society. This comprehensive integration opens new horizons for data-driven healthcare advancements, personalized medicine, and evidence-based decision-making. Responsible integration of clinical trial data into EHDS and TEHDAS with dataspace and machine learning sets a strong foundation for transformative research that drives positive patient outcomes, advances medical knowledge, and ultimately improves healthcare practices in Europe and beyond.

VI. Future Directions

The future direction of integrating linked graph dataspace, machine learning, and clinical trials holds significant promise for advancing healthcare research and treatment strategies. As technology continues to evolve, there is a growing emphasis on refining and expanding the methodologies applied to clinical trial data analysis. Enhanced data integration techniques, fueled by semantic engines and knowledge graph embedding algorithms, are anticipated to enable richer insights by capturing intricate relationships within the data.

Further advancements in graph-based machine learning approaches are likely to enhance predictive modeling accuracy, enabling more accurate patient outcome predictions. As datasets expand and become more complex, the development of scalable and efficient algorithms becomes paramount, allowing researchers to effectively navigate and extract insights from the vast linked graph dataspace. Ethical considerations and regulatory frameworks will play an increasingly crucial role in shaping the future of these endeavors. Striking a balance between data sharing for research purposes and patient privacy protection will remain a central challenge.

Moreover, as healthcare systems transition towards more patient-centric approaches, the integration of personalized medicine principles within linked graph dataspace holds immense potential. Tailoring treatment strategies to individual patients based on comprehensive linked graph analyses could revolutionize clinical decision-making and patient care. Collaboration across multidisciplinary teams, including data scientists, clinicians, regulatory experts, and patients, will be key to harnessing the full potential of linked graph dataspace, machine learning, and clinical trials. Continued investment in

[4]https://gdpr.eu/what-is-gdpr/

[5]https://www.european-health-data-space.com/

[6]https://tehdas.eu/

research, technology development, and infrastructure will pave the way for a future where healthcare is optimized through innovative, data-driven approaches.

VII. Conclusion

In conclusion, the convergence of linked graph dataspace, machine learning, and clinical trials represents a transformative paradigm in healthcare research. The seamless integration of diverse data sources within a structured graph framework promises to unlock valuable insights and drive evidence-based medical advancements. By leveraging graph-based machine learning, researchers can unearth hidden patterns and relationships that were previously elusive, leading to more accurate predictive models and personalized treatment strategies. While this synergy holds immense potential, it also presents formidable challenges. Ethical considerations and regulatory frameworks must be carefully navigated to ensure patient privacy and data security. The complexity of integrating heterogeneous data sources and maintaining data quality demands innovative solutions and robust algorithms. As the healthcare landscape continues to evolve, the future direction of this field holds great promise. From enhancing patient outcomes through personalized medicine to advancing clinical trial methodologies, the integration of linked graph dataspace and machine learning stands as a beacon of progress. Collaborative efforts across various disciplines will drive innovation and transform healthcare into a more data-driven and patient-centric domain. In the coming years, as technology advances and collaborations flourish, the vision of harnessing the collective power of linked graph dataspace, machine learning, and clinical trials will materialize into tangible benefits for patients, clinicians, researchers, and the broader healthcare ecosystem. Through persistent exploration, refinement, and ethical stewardship, this paradigm could shape the future of healthcare.

Acknowledgment

We would like to acknowledge Science Foundation Ireland (SFI/12/RC/2289_P2) for funding this research.

References

[1] M. Hildebrandt, "Ground-truthing in the european health data space," *SocArXiv. January*, vol. 12, 2023.

[2] V. Huser and J. J. Cimino, "Impending challenges for the use of big data," *International journal of radiation oncology, biology, physics*, vol. 95, no. 3, pp. 890–894, 2016.

[3] A. C. Machado and D. F. Polónia, "Legal and technological aspects for the creation of a european health data space," in *2022 17th Iberian Conference on Information Systems and Technologies (CISTI)*. IEEE, 2022, pp. 1–6.

[4] E. J. Topol, "High-performance medicine: the convergence of human and artificial intelligence," *Nature medicine*, vol. 25, no. 1, pp. 44–56, 2019.

[5] S. K. Bansal, "Towards a semantic extract-transform-load (etl) framework for big data integration," in *2014 IEEE International Congress on Big Data*. IEEE, 2014, pp. 522–529.

[6] R. Reda, F. Piccinini, G. Martinelli, and A. Carbonaro, "Heterogeneous self-tracked health and fitness data integration and sharing according to a linked open data approach," *Computing*, vol. 104, no. 4, pp. 835–857, 2022.

[7] A. Valls, K. Gibert, D. Sánchez, and M. Batet, "Using ontologies for structuring organizational knowledge in home care assistance," *international journal of medical informatics*, vol. 79, no. 5, pp. 370–387, 2010.

[8] Q. Guo, C. Zhang, S. Zhang, and J. Lu, "Multi-model query languages: taming the variety of big data," *Distributed and Parallel Databases*, pp. 1–41, 2023.

[9] M. G. Skjæveland, K. Balog, N. Bernard, W. Lajewska, and T. Linjordet, "An ecosystem for personal knowledge graphs: A survey and research roadmap," *arXiv preprint arXiv:2304.09572*, 2023.

[10] K. Verma and A. Kass, "Requirements analysis tool: A tool for automatically analyzing software requirements documents," in *The Semantic Web-ISWC 2008: 7th International Semantic Web Conference, ISWC 2008, Karlsruhe, Germany, October 26-30, 2008. Proceedings 7*. Springer, 2008, pp. 751–763.

[11] J. Cardoso and A. Sheth, "The semantic web and its applications," in *Semantic Web Services, Processes and Applications*. Springer, 2006, pp. 3–33.

[12] G. Laleci, G. Aluc, A. Dogac, A. Sinaci, O. Kilic, and F. Tuncer, "A semantic backend for content management systems," *Knowledge-based systems*, vol. 23, no. 8, pp. 832–843, 2010.

[13] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," *arXiv preprint arXiv:1609.02907*, 2016.

[14] F. Scarselli, M. Gori, A. C. Tsoi, M. Hagenbuchner, and G. Monfardini, "The graph neural network model," *IEEE transactions on neural networks*, vol. 20, no. 1, pp. 61–80, 2008.

[15] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Lio, and Y. Bengio, "Graph attention networks," *arXiv preprint arXiv:1710.10903*, 2017.

[16] A. Janik, M. Torrente, L. Costabello, V. Calvo, B. Walsh, C. Camps, S. K. Mohamed, A. L. Ortega, V. Nováček, B. Massutí *et al.*, "Machine learning–assisted recurrence prediction for patients with early-stage non–small-cell lung cancer," *JCO Clinical Cancer Informatics*, vol. 7, p. e2200062, 2023.

[17] T. Lacroix, N. Usunier, and G. Obozinski, "Canonical tensor decomposition for knowledge base completion," in *International Conference on Machine Learning*. PMLR, 2018, pp. 2863–2872.

[18] D. A. Grimes and K. F. Schulz, "An overview of clinical research: the lay of the land," *The lancet*, vol. 359, no. 9300, pp. 57–61, 2002.

[19] A. Goldberg, L. N. Bakhireva, K. Page, and A. M. Henrie, "A qualitative scoping review of early-terminated clinical trials sponsored by the department of veterans affairs cooperative studies program from 2010 to 2020," *Epidemiologic Reviews*, vol. 44, no. 1, pp. 110–120, 2022.

[20] E. H. Weissler, T. Naumann, T. Andersson, R. Ranganath, O. Elemento, Y. Luo, D. F. Freitag, J. Benoit, M. C. Hughes, F. Khan *et al.*, "The role of machine learning in clinical research: transforming the future of evidence generation," *Trials*, vol. 22, no. 1, pp. 1–15, 2021.

[21] F. D. Beacher, L. R. Mujica-Parodi, S. Gupta, and L. A. Ancora, "Machine learning predicts outcomes of phase iii clinical trials for prostate cancer," *Algorithms*, vol. 14, no. 5, p. 147, 2021.

[22] K. M. Gayvert, N. S. Madhukar, and O. Elemento, "A data-driven approach to predicting successes and failures of clinical trials," *Cell chemical biology*, vol. 23, no. 10, pp. 1294–1301, 2016.

[23] E. Kavalci and A. Hartshorn, "Improving clinical trial design using interpretable machine learning based prediction of early trial termination," *Scientific Reports*, vol. 13, no. 1, p. 121, 2023.

[24] M. Timilsina, M. Tandan, M. d'Aquin, and H. Yang, "Discovering links between side effects and drugs using a diffusion based method," *Scientific reports*, vol. 9, no. 1, p. 10436, 2019.

[25] M. Timilsina, M. Tandan, and V. Nováček, "Machine learning approaches for predicting the onset time of the adverse drug events in oncology," *Machine Learning with Applications*, vol. 9, p. 100367, 2022.

[26] M. Timilsina, D. Fey, S. Buosi, A. Janik, L. Costabello, E. Carcereny, D. R. Abreu, M. Cobo, R. L. Castro, R. Bernabé *et al.*, "Synergy between imputed genetic pathway and clinical information for predicting recurrence in early stage non-small cell lung cancer," *Journal of Biomedical Informatics*, p. 104424, 2023.

[27] J. Zhao, A. Henriksson, L. Asker, and H. Boström, "Predictive modeling of structured electronic health records for adverse drug event detection," *BMC medical informatics and decision making*, vol. 15, no. 4, pp. 1–15, 2015.

[28] D. Horgan, M. Hajduch, M. Vrana, J. Soderberg, N. Hughes, M. I. Omar, J. A. Lal, M. Kozaric, F. Cascini, V. Thaler *et al.*, "European health data space—an opportunity now to grasp the future of data-driven healthcare," in *Healthcare*, vol. 10, no. 9. MDPI, 2022, p. 1629.

[29] A. Gyrard, M. Gaur, S. Shekarpour, K. Thirunarayan, and A. Sheth, "Personalized health knowledge graph," in *CEUR workshop proceedings*, vol. 2317. NIH Public Access, 2018.

[30] L. Li, P. Wang, J. Yan, Y. Wang, S. Li, J. Jiang, Z. Sun, B. Tang, T.-H. Chang, S. Wang *et al.*, "Real-world data medical knowledge graph: construction and applications," *Artificial intelligence in medicine*, vol. 103, p. 101817, 2020.

[31] B. A. Badwan, G. Liaropoulos, E. Kyrodimos, D. Skaltsas, A. Tsirigos, and V. G. Gorgoulis, "Machine learning approaches to predict drug efficacy and toxicity in oncology," *Cell Reports Methods*, vol. 3, no. 2, 2023.

[32] M. Yagi, K. Yamanouchi, N. Fujita, H. Funao, and S. Ebata, "Revolutionizing spinal care: Current applications and future directions of artificial intelligence and machine learning," *Journal of Clinical Medicine*, vol. 12, no. 13, p. 4188, 2023.

[33] S. Paredes, J. Henriques, T. Rochar, D. Mendes, P. Carvalho, J. Moraisl, A. Bianchi, and V. Salcedof, "A clinical interpretable approach applied to cardiovascular risk assessment," in *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2018, pp. 3252–3255.

[34] M. Tandan, M. Timilsina, M. Cormican, and A. Vellinga, "Role of patient descriptors in predicting antimicrobial resistance in urinary tract infections using a decision tree approach: a retrospective cohort study," *International journal of medical informatics*, vol. 127, pp. 127–133, 2019.

[35] L. Jiang, L. Li, H. Cai, H. Liu, J. Hu, and C. Xie, "A linked data-based approach for clinical treatment selecting support," *Journal of Management Analytics*, vol. 1, no. 4, pp. 301–316, 2014.

[36] M. Timilsina, S. Buosi, D. Fey, A. Janik, M. Torrente, M. Provencio, A. C. Bermu, E. Carcereny, L. Costabello, D. Rodr *et al.*, "Integration of clinical information and imputed aneuploidy scores to enhance relapse prediction in early stage lung cancer patients," in *AMIA Annual Symposium Proceedings*, vol. 2022. American Medical Informatics Association, 2022, p. 1062.

[37] M. Timilsina, H. Yang, R. Sahay, and D. Rebholz-Schuhmann, "Predicting links between tumor samples and genes using 2-layered graph based diffusion approach," *BMC bioinformatics*, vol. 20, no. 1, pp. 1–20, 2019.

[38] K. B. Johnson, W.-Q. Wei, D. Weeraratne, M. E. Frisse, K. Misulis, K. Rhee, J. Zhao, and J. L. Snowdon, "Precision medicine, ai, and the future of personalized health care," *Clinical and translational science*, vol. 14, no. 1, pp. 86–93, 2021.

[39] J. Yang, Y. Li, Q. Liu, L. Li, A. Feng, T. Wang, S. Zheng, A. Xu, and J. Lyu, "Brief introduction of medical database and data mining technology in big data era," *Journal of Evidence-Based Medicine*, vol. 13, no. 1, pp. 57–69, 2020.

[40] A. Winter, S. Stäubert, D. Ammon, S. Aiche, O. Beyan, V. Bischoff, P. Daumke, S. Decker, G. Funkat, J. E. Gewehr *et al.*, "Smart medical information technology for healthcare (smith)," *Methods of information in medicine*, vol. 57, no. S 01, pp. e92–e105, 2018.

[41] A. Qayyum, J. Qadir, M. Bilal, and A. Al-Fuqaha, "Secure and robust machine learning for healthcare: A survey," *IEEE Reviews in Biomedical Engineering*, vol. 14, pp. 156–180, 2020.

[42] O. P. Jena, B. Bhushan, N. Rakesh, P. N. Astya, and Y. Farhaoui, *Machine Learning and Deep Learning in Efficacy Improvement of Healthcare Systems*. CRC Press, 2022.

[43] K. M. Boehm, E. A. Aherne, L. Ellenson, I. Nikolovski, M. Alghamdi, I. Vázquez-García, D. Zamarin, K. Long Roche, Y. Liu, D. Patel *et al.*, "Multimodal data integration using machine learning improves risk stratification of high-grade serous ovarian cancer," *Nature cancer*, vol. 3, no. 6, pp. 723–733, 2022.

[44] C. Timmermans, D. Venet, and T. Burzykowski, "Data-driven risk identification in phase iii clinical trials using central statistical monitoring," *International journal of clinical oncology*, vol. 21, pp. 38–45, 2016.

[45] M. Wang, S. Li, T. Zheng, N. Li, Q. Shi, X. Zhuo, R. Ding, Y. Huang *et al.*, "Big data health care platform with multisource heterogeneous data integration and massive high-dimensional data governance for large hospitals: Design, development, and application," *JMIR Medical Informatics*, vol. 10, no. 4, p. e36481, 2022.

[46] S. Sanchez-Martinez, O. Camara, G. Piella, M. Cikes, M. Á. González-Ballester, M. Miron, A. Vellido, E. Gómez, A. G. Fraser, and B. Bijnens, "Machine learning for clinical decision-making: challenges and opportunities in cardiovascular imaging," *Frontiers in Cardiovascular Medicine*, vol. 8, p. 765693, 2022.

[47] C. Selvaraj, I. Chandra, and S. K. Singh, "Artificial intelligence and machine learning approaches for drug design: challenges and opportunities for the pharmaceutical industries," *Molecular diversity*, pp. 1–21, 2021.

[48] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Artificial intelligence and statistics*. PMLR, 2017, pp. 1273–1282.

[49] M. A. Krahe, M. Wolski, S. Mickan, J. Toohey, P. Scuffham, and S. Reilly, "Developing a strategy to improve data sharing in health research: A mixed-methods study to identify barriers and facilitators," *Health Information Management Journal*, vol. 52, no. 1, pp. 18–27, 2023.

[50] A. Torab-Miandoab, T. Samad-Soltani, A. Jodati, and P. Rezaei-Hachesu, "Interoperability of heterogeneous health information systems: a systematic literature review," *BMC Medical Informatics and Decision Making*, vol. 23, no. 1, p. 18, 2023.

[51] N. Rieke, J. Hancox, W. Li, F. Milletari, H. R. Roth, S. Albarqouni, S. Bakas, M. N. Galtier, B. A. Landman, K. Maier-Hein *et al.*, "The future of digital health with federated learning," *NPJ digital medicine*, vol. 3, no. 1, p. 119, 2020.

[52] G. A. Kaissis, M. R. Makowski, D. Rückert, and R. F. Braren, "Secure, privacy-preserving and federated machine learning in medical imaging," *Nature Machine Intelligence*, vol. 2, no. 6, pp. 305–311, 2020.

[53] H. Hallock, S. E. Marshall, P. A. C. 't Hoen, J. F. Nygård, B. Hoorne, C. Fox, and S. Alagaratnam, "Federated networks for distributed analysis of health data," *Frontiers in Public Health*, vol. 9, 2021. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fpubh.2021.712569

[54] A. Hasnain, M. R. Kamdar, P. Hasapis, D. Zeginis, C. N. Warren, H. F. Deus, D. Ntalaperas, K. Tarabanis, M. Mehdi, and S. Decker, "Linked biomedical dataspace: lessons learned integrating data for drug discovery," in *The Semantic Web–ISWC 2014: 13th International Semantic Web Conference, Riva del Garda, Italy, October 19-23, 2014. Proceedings, Part I 13*. Springer, 2014, pp. 114–130.

[55] Y. S. Can and C. Ersoy, "Privacy-preserving federated deep learning for wearable iot-based biomedical monitoring," *ACM Transactions on Internet Technology (TOIT)*, vol. 21, no. 1, pp. 1–17, 2021.

[56] M. Ghassemi, L. Oakden-Rayner, and A. L. Beam, "The false hope of current approaches to explainable artificial intelligence in health care," *The Lancet Digital Health*, vol. 3, no. 11, pp. e745–e750, 2021.