THERAPEUTIC ADVANCES in
*Drug Safety*

*Review*

# Perspective review: Will generative AI make common data models obsolete in future analyses of distributed data networks?

*Ther Adv Drug Saf*

2025, Vol. 16: 1–17

DOI: 10.1177/20420986251332743



**Jeffery L. Painter, Darmendra Ramcharran and Andrew Bate**

***Abstract*:** Integrating real-world healthcare data is challenging due to diverse formats and terminologies, making standardization resource-intensive. While Common Data Models (CDMs) facilitate interoperability, they often cause information loss, exhibit semantic inconsistencies, and are labor-intensive to implement and update. We explore how generative artificial intelligence (GenAI), especially large language models (LLMs), could make CDMs obsolete in quantitative healthcare data analysis by interpreting natural language queries and generating code, enabling direct interaction with raw data. Knowledge graphs (KGs) standardize relationships and semantics across heterogeneous data, preserving integrity. This perspective review proposes a fourth generation of distributed data network analysis, building on previous generations categorized by their approach to data standardization and utilization. It emphasizes the potential of GenAI to overcome the limitations CDMs with GenAI-enabled access, KGs, and automatic code generation. A data commons may further enhance this capability, and KGs may well be needed to enable effective GenAI. Addressing privacy, security, and governance is critical; any new method must ensure protections comparable to CDM-based models. Our approach would aim to enable efficient, real-time analyses across diverse datasets and enhance patient safety. We recommend prioritizing research to assess how GenAI can transform quantitative healthcare data analysis by overcoming current limitations.

## Plain language summary

### Perspective Review: Will generative AI make common data models obsolete in future analyses of distributed data networks?

This perspective review explores whether Artificial Intelligence (AI) can revolutionize healthcare data analysis by reducing the current reliance on Common Data Models (CDMs), which encompass the following elements:

- CDMs are approaches that standardize diverse healthcare to a single shared format to enable efficiencies in data management and analyses using the same analysis syntax and analytic tools.
- Although CDMs have strengths, they also have limitations, such as high costs, potential loss of important details, significant effort to produce and maintain, and delays in data availability due to lengthy data processing steps.
- With the rapid growth of healthcare data, effectively analyzing it is crucial for patient safety and public health.
- AI may offer an alternative solution by analyzing data directly in its original form, reducing costs, preserving data details, and enabling real-time insights that support better patient outcomes and safer medication use.

Correspondence to:
**Jeffery L. Painter**
GSK, 410 Blackwell Street, Durham, NC 27701, USA
**jeffery.l.painter@gsk.com**

**Darmendra Ramcharran**
GSK, Providence, RI, USA

**Andrew Bate**
GSK, London, UK

London School of Hygiene and Tropical Medicine, London, UK

This review investigates challenges currently associated with CDMs and explores how AI, particularly generative AI, can directly analyze raw data without the need for standardization. We discuss the following:

- How AI can interpret complex questions and generate accurate answers from raw data, enabling more timely analyses of real-world data.
- While CDMs may still be necessary in the short term, AI has the potential to eventually replace them, improving patient care and safety outcomes by providing faster and more precise insights.
- This perspective could lead to new methods of using healthcare data to inform decision-making and enhance treatment outcomes.
- By adopting advanced AI technologies, healthcare providers and researchers can better understand treatment risks and benefits, make more informed decisions, and ultimately improve patient safety and public health.

***Keywords:*** Common Data Model (CDM), distributed data network, drug safety, generative AI, pharmacovigilance

Received: 3 October 2024; revised manuscript accepted: 19 March 2025.

## Introduction

This manuscript represents a perspective review, synthesizing insights from literature and expert consensus to evaluate the evolving role of Common Data Models (CDMs) in distributed analytics.

Over the past several decades, electronic healthcare databases have evolved from simple medical record repositories into sophisticated tools for complex epidemiological research, informing regulatory, clinical, and policy decisions.[1] Alongside randomized controlled trials, registries, and spontaneous reports, they are pivotal for understanding and ensuring medication safety during drug development and clinical use.

These transformations can largely be categorized into generational developments that we have defined as follows (Figure 1). This generational framework is based on the authors' synthesis of advancements in healthcare data analysis and their implications for real-world evidence generation. The first generation (initiated circa 1970) featured proprietary systems relying on single-database analyses without standardized formats.[2–4] The second generation (initiated circa 1991) introduced widely used coding standards and enterprise databases,[5] but data sharing involved individual contracts and bespoke cross-database comparisons or substantial computational resources for meta-analyses across multiple proprietary sources.[6,7]

The third generation, starting in the late 1990s and gaining momentum after the FDA Amendments Act of 2007 (https://www.fda.gov/regulatory-information/selected-amendments-fdc-act/food-and-drug-administration-amendments-act-fdaaa-2007) and the launch of Mini-Sentinel in 2009,[8] is characterized by the adoption of CDMs.[9] Early CDMs like the Vaccine Safety Datalink's (VSD's)[10,11] demonstrated the feasibility of CDMs. While the Observational Medical Outcomes Partnership's (OMOP's) CDM (https://www.ohdsi.org/data-standardization/) popularized distributed data networks (DDNs) and open-source methodologies, greatly improving interoperability and streamlining analyses across varied databases.[5,8,12,13] Once a database was converted to a given CDM external standardized analyses could be conducted with the agreement of the database holder across the database with the need to only share summarized results. Similarly, standardized

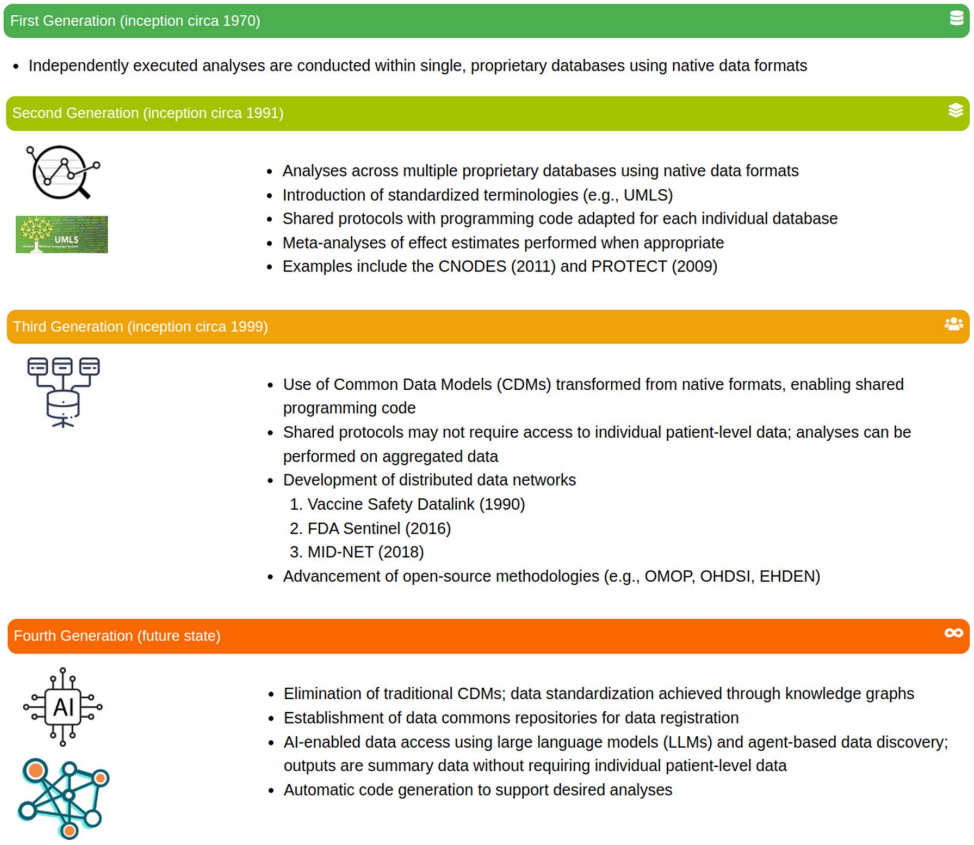


**Figure 1.** Four generations of quantitative analysis in healthcare data.

tools and programs developed for the CDM could be readily used by the database holder on their own data.

However, despite their significant contributions, CDMs are not without limitations, particularly when scaling analyses across increasingly complex and heterogeneous data sources. The advent of generative artificial intelligence (GenAI) and knowledge graphs (KGs) heralds a potential fourth generation in electronic healthcare data utilization. These technologies promise to integrate, analyze, and interpret multiple datasets—whether for distributed analyses, multi-site studies, or other complex use cases—without requiring a single standardized format. KGs are an evolving research tool and will enable increasingly effective use of GenAI in enabling analyses across different data formats.

This paper explores the relevance of CDMs considering these emerging innovations, examining whether GenAI can address current challenges and render traditional CDMs less central to future healthcare data analyses.

## Background

### *The role of CDMs*

*History and development of DDNs and CDM utilization.* The development of DDNs has evolved through two primary approaches: the common protocol approach and the CDM approach. However, sometimes a hybrid of the two approaches was used where a study-specific CDM was implemented (e.g., earlier studies in the Asian Pharmacoepidemiology Network (AsPEN)). Early networks, such as the Canadian Network for Observational Drug Effect Studies (CNODES; https://www.cnodes.ca/), and the Pharmacoepidemiological Research on Outcomes of Therapeutics by a European Consortium (PROTECT; https://www.imi.europa.eu/projects-results/project-factsheets/protect), adopted the common protocol approach. This method allowed each

participating site to retain data in its original format while following standardized study protocols and centrally developed code.[14,15] While this approach enabled data custodians to maintain control and minimized the transformation workload, it also introduced significant challenges in data standardization. Harmonizing data across sites proved time-intensive, increasing the complexity of cross-database analyses and limiting scalability, particularly in scenarios requiring rapid responses.[16]

To address these limitations, the CDM approach was introduced to standardize data structures across multiple sites. By transforming data into a uniform format before analysis, CDMs aimed to enhance comparability and efficiency in multi-database research. However, the adoption of CDMs also presented trade-offs, including resource-intensive data transformation processes, potential information loss, and challenges in achieving consistency across different CDMs.[17] These trade-offs highlight the ongoing need for innovative solutions to bridge gaps in interoperability and efficiency within DDNs.

*Global perspectives on CDM usage.* Internationally, various federated DDNs use CDMs to enhance health-related data analyses. An early example is the VSD project,[10,11] which established a CDM to enable secure, federated querying across multiple health maintenance organizations. This model facilitated vaccine safety surveillance while addressing concerns about data privacy and confidentiality. Additional examples include the Medical Information Database Network (MIDNET),[18] the AsPEN (https://www.aspensig.asia/), TriNetX,[19] and CNODES.[14]

Several other global networks have adopted CDMs to standardize health data integration. Networks like the OMOP and the Observational Health Data Sciences and Informatics (OHDSI; https://www.ohdsi.org/) program promote a highly structured CDM to support cross-network interoperability.[20] Others, such as the Sentinel Initiative (https://www.sentinelinitiative.org/), NorPEN (https://www.norpen.org/), IMI ConcePTION (https://www.imi-conception.eu/), PCORnet (https://pcornet.org/), and the DARWIN-EU (https://www.darwin-eu.org), European Health Data and Evidence Network (EHDEN; https://www.ehden.eu/), employ different CDMs tailored to specific regulatory and research needs.[21] Meanwhile, initiatives like DARWIN-EU facilitate extensive data integration across diverse European sources using the OMOP CDM.

Despite the advantages of standardization, concerns remain regarding information loss when mapping diverse datasets to a standardized model. These concerns drive ongoing refinement of CDMs to balance standardization with the need for preserving the granularity and semantic integrity of source data.[22,23] Alternative approaches, such as the Generalized Data Model (GDM), have been proposed to maintain data in their native formats while leveraging clinical codes and hierarchical structures.[22] However, while these alternatives mitigate information loss, they also introduce complexities in data analysis and may extend the timeline required to generate insights.

### *Challenges in CDM implementation*

The adoption and implementation of CDMs present several challenges, as identified through literature reviews and expert discussions in DDN analysis. The growing size and diversity of data networks intensify the challenges of converting data to a standard CDM, compounded by the constantly changing healthcare landscape.[24] Integrating heterogeneous data sources complicates the creation of universally applicable CDMs essential for drug safety surveillance.[25] Ensuring semantic consistency is crucial to avoid misinterpretations in drug safety analysis.

Kent et al. highlight the difficulties in standardizing data across healthcare systems, where variability in data collection methods can compromise analysis reliability.[26] Overcoming technological, methodological, regulatory, and ethical challenges adds complexity to CDM implementation. Analysts also struggle with accessing data through various vendor tools and employing different strategies, which complicates analyses and prolongs insight generation.[27] In addition, evolving CDMs and their ecosystems can lead to discrepancies in outputs, further complicating comparisons across different models.[3,28,29] Other CDM implementation challenges, whether using the pragmatic or generic approach, are defined as follows:

1. *Information loss and data semantics:* Transitioning data into a CDM can lead to substantial information loss, especially if

the CDM does not perfectly represent the source data.[30–34] Garza et al. emphasized that unless a CDM perfectly represents the source data, information loss will occur, potentially altering the original data semantics.[35] This issue is pronounced with registries, which are heterogeneous in data collection and structure.[36]

2. *Resource intensity of data transformation:* Transforming and maintaining data in a CDM is resource-intensive, often deterring organizations with limited resources.[22] Frequent updates and ensuring compatibility across different CDM versions add to the complexity and workload.[34]
3. *Variability in data sources:* Standardization is challenging due to variability in healthcare delivery and data capture across regions and time, risking the loss of critical data nuances.[22] Unclear assumptions during data conversion can lead to overconfident decisions based on flawed interpretations, underscoring the need for transparent documentation and critical evaluation of the transformation process.
4. *Diverse CDMs and analytical tools:* Variations in CDMs and their associated analytical ecosystems can yield differing results, complicating decision-making.[28] Careful selection of a particular model and its tools is essential, as different models may impact the generated evidence.

These challenges have prompted the exploration of alternative approaches to data integration and standardization. One such alternative is the GDM proposed by Danese et al..[22] The GDM focuses on retaining the original semantic representation using clinical codes in their native vocabularies and preserving hierarchical information and provenance. By avoiding transformation into a standardized CDM, the GDM aims to reduce information loss and maintain data integrity.

However, while the GDM addresses some challenges of traditional CDMs, it introduces its own limitations. Analyses using the GDM can be time-consuming, taking longer from ideation to execution.[22] This extended timeline may impact the timeliness of research findings, especially in fast-moving fields where rapid insights are critical. The complexity of working with heterogeneous data in native formats may require more sophisticated analytical tools and expertise, potentially limiting accessibility for some organizations.

*Challenges with terminologies and ontologies in real-world data networks.* In real-world data (RWD) networks, integrating and analyzing data from diverse sources is often complicated by variations in medical terminologies and ontologies. Common terminologies used historically in healthcare include coding systems such as the International Classification of Diseases (e.g., ICD-9 and ICD-10), Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT), and Logical Observation Identifiers Names and Codes (LOINC). Many other terminologies are employed based on specific subject areas, such as lab values (e.g., LOINC for laboratory tests) and billing codes (e.g., Current Procedural Terminology). These terminologies are essential for coding diagnoses, procedures, and other clinical information within electronic health and medical records.[37] However, the diversity of these coding systems necessitates complex mapping and cross-referencing to achieve consistent data integration, posing significant challenges for data harmonization and interoperability in RWD networks.

*Variability of terminologies across systems:* Different healthcare systems and regions may use different coding systems or versions of the same system. For instance, while some systems use ICD-9-CM, others have transitioned to ICD-10-CM, and still others may use SNOMED CT or Read codes.[38] This variability poses significant challenges for data integration and interoperability within RWD networks.

*Challenges in mapping between terminologies:* Cross-mapping between different coding systems is necessary for integrating data from multiple sources. However, these mappings can be complex and may not capture the full semantic relationships between codes, leading to potential information loss or misinterpretation.[39] For example, mapping from ICD codes to SNOMED CT or Medical Dictionary for Regulatory Activities (MedDRA) may not always be straightforward due to differences in granularity and coding structures.[40]

*MedDRA and its integration with RWD terminologies:* In drug safety surveillance, the MedDRA is the standard terminology used for coding adverse

events.[41] Integrating MedDRA-coded data with RWD, which often uses ICD codes or other terminologies, requires accurate and reliable mapping. Inconsistencies or inaccuracies in these mappings can affect the identification and evaluation of safety signals.[42]

*Versioning and updates:* Frequent updates and version changes in coding systems add another layer of complexity. Ensuring that mappings remain accurate over time requires continuous maintenance and updates, which can be resource-intensive.[43]

*Leveraging AI and KGs:* AI and KGs offer promising solutions to these challenges. AI techniques can automate the mapping between different terminologies, enhancing accuracy and reducing manual effort.[44] KGs can represent complex relationships between medical concepts across different coding systems, facilitating interoperability and integration.[45]

As we anticipate an increasing need for more multimodal quantitative surveillance, the ability to work rapidly and effectively across diverse data streams becomes ever more important. Therefore, standards applied to the collection and reporting of Individual Case Safety Reports and MedDRA, along with their links to RWD terminologies like ICD codes, are essential for ensuring consistency and reliability in data integration and analysis.[46]

By addressing these challenges through the adoption of AI-driven solutions and adherence to standardized terminologies, RWD networks can improve data interoperability and enhance the quality of healthcare analytics, ultimately contributing to better patient outcomes and more informed public health decisions.

### *Additional challenges faced in multi-database analyses*

Multi-database analyses leverage data from various sources, offering significant benefits for rapid-cycle analyses and safety signal detection in healthcare. However, their widespread adoption faces challenges for data analysts, custodians, and stakeholders who require rapid, actionable, and trustworthy outputs to impact public health and patient care.

*The continual search for fit-for-purpose RWD.* Identifying ideal RWD sources that accurately represent target populations is inherently challenging. Each database captures only a subset of the population under specific healthcare settings, introducing biases related to demographics, geography, and healthcare access. Data must be fit for specific study purposes and sufficiently recent to avoid misleading insights about current practices and outcomes.[47] Researchers must select data sources closely aligned with their objectives, carefully weighing strengths and limitations to minimize bias.

For data custodians, maintaining databases in a CDM or multiple CDMs poses a significant workload, especially if the CDM does not align with their core database model. Adopting a CDM that does not perfectly map to the primary data use may require compromises on data utility or adopting a simpler CDM closer to specific databases. This challenge explains why new CDMs continue to be proposed despite long-standing models like Sentinel and OMOP .[22,48]

Ensuring trustworthiness requires considering how extensively databases and their conversions need to be documented and preserved for future reanalysis. For data providers, extensive documentation of transformations can be burdensome. While replication—achieving consistent results using the same methods on similar data—can suffice,[49] stakeholders may demand reproducibility to ensure validity,[50] requiring detailed documentation and preservation, adding to custodians' workload.

Trustworthiness also depends on producing rapid, actionable outputs impacting public health and patient care. Data custodians must balance timely data updates with maintaining CDMs, which may not align with their primary operations. This tension underscores the importance of developing more adaptable data models or alternative approaches that reduce providers' workload while ensuring data remains fit-for-purpose.

*Privacy concerns and data sharing complexities.* Privacy regulations, such as the General Data Protection Regulation[51] in the European Union and the Health Insurance Portability and Accountability Act[52] in the United States, impose stringent controls on how personal health

information (PHI) is collected, stored, and shared. While protecting patient confidentiality, these regulations complicate data sharing, often causing delays and increased costs.[53] Data custodians must invest substantial resources to ensure compliance, limiting their ability to share data promptly.

Logistical complexities in data sharing, especially internationally, are considerable. Differing regulations and standards across countries complicate the harmonization of data collection and analysis procedures,[15,54] necessitating sophisticated data governance structures. For stakeholders requiring rapid and trustworthy outputs, these delays can hinder timely decision-making for public health and patient care.

Commercial concerns and data format limitations further restrict data availability, as highlighted by Walker et al.[55] Privacy concerns particularly affect access to PHI, such as unstructured clinical notes, limiting the depth of possible analysis.

*Decentralization versus centralization of data.* To balance comprehensive analysis with privacy concerns, initiatives like the FDA's Sentinel network maintain decentralized patient-level data while centralizing summary data for analysis.[53] This model permits detailed analyses locally, sharing only aggregated results centrally. For data custodians, this allows control over sensitive data, ensuring compliance and preserving primary database purposes.

However, decentralization can restrict the depth and speed of analysis stakeholders desire. Sensitive data remains at the source, potentially slowing the generation of rapid, actionable insights needed for public health decisions and patient care. Stakeholders must rely on custodians' capacity and willingness to perform analyses promptly and accurately. Furthermore, clear governance frameworks and secure access protocols are essential to ensure sensitive data are only used and shared in compliance with regulatory and ethical standards.

*Technical challenges in data aggregation and synthesis.* Combining data from multiple databases poses significant technical challenges. Variations in data capture methodologies, variable definitions, coding systems, and missing data impact the quality and comparability of the aggregated dataset. Researchers must utilize sophisticated harmonization techniques—often requiring complex algorithms and substantial computational resources—to ensure the combined data accurately reflects underlying realities.

For data custodians, the workload to keep a database updated in a CDM or different CDMs is considerable, especially if these models are not integral to their primary operations. Adopting a CDM not perfectly aligned with primary data use can necessitate compromises or additional resources. This is why new CDMs continue to be proposed, aiming for closer alignment with specific datasets.[48]

For stakeholders seeking rapid and trustworthy outputs, technical challenges in data aggregation can cause delays and affect reliability. Ensuring data transformations and analyses are well-documented and reproducible enhances trust but adds complexity and time to deliver actionable insights.

Despite these challenges, data custodians recognize two primary benefits of converting to CDMs: (1) the ability to use tools developed within an ecosystem designed for a given CDM and (2) the capability to readily participate in multiple database studies based on that common structure. However, we believe that in an AI-enabled future, these benefits may be more easily attainable without relying on traditional CDMs. Advanced AI techniques could facilitate data integration and analysis across heterogeneous databases, reducing custodians' burden and accelerating the delivery of actionable insights to stakeholders.

## Discussion

Having examined the evolution of the CDM and its current role in PV leads us to wonder what comes next. As we alluded to in the introduction, we believe there is a strong role to play for GenAI and KGs to enable the fourth generation of safety surveillance. Below, we will explore the building blocks necessary to move us into the fourth generation, as well as explore potential challenges and considerations the PV community must face as we move into the next phase of technology to assist in PV-related activities.

### *Building blocks of the fourth generation enabled safety surveillance*

We propose a fourth generation of AI-enabled data analysis to simplify multi-database analysis and ease data availability for custodians. Key elements may include the following: (1) establishing data commons for data registration, (2) utilizing AI-enabled data access via agent-based discovery, (3) standardizing data through KGs, (4) supporting automatic code generation for analyses, and (5) managing operational aspects like master service agreements to protect personally identifiable information, facilitate reanalyses, and ease access while respecting data localization laws and collaboration requirements. This approach could reduce the need to transform native data into CDMs, enabling efficient, real-time analysis across diverse datasets. To realize this potential, we recommend prioritizing research that assesses how GenAI can accelerate the effective use of electronic health data to enhance patient safety.

The evolution of drug safety surveillance into its fourth generation requires a paradigm shift toward more dynamic and interconnected data ecosystems. Central to this shift is the ability to seamlessly integrate and analyze data from diverse sources—whether in-house or external. This integration first relies on the concept of data "discoverability," ensuring that data systems are easily findable and comprehensively described by metadata that details their contents and relevance to specific studies. Notably, although ontologies for RWD metadata are important to enable this paradigm shift, the extent to which these metadata exist and are available is insufficient. Processes, potentially supported with GenAI tools, may support the development and make available ontologies for RWD metadata, as are financial support and a standardized approach to ensure metadata maintenance and updating. Recent frameworks like the DIVERSE framework and the MINERVA metadata list offer promising strategies for standardizing metadata across diverse data sources.[56,57] Such tools are factors that may support interoperability and reduce bias, particularly in distributed analytics.

DDNs enhance statistical power by aggregating data from multiple sources, enabling robust analyses that can detect rare adverse events and improve generalizability across diverse populations and settings.[58]

However, integrating varied data sources presents challenges, especially for pharmacovigilance (PV). Safety analyses often use data not optimized for population-level studies, leading to issues in data availability and technical barriers due to different formats and coding systems, which can introduce selection bias.[59] The Structured Process to Identify Fit-For-Purpose Data (SPIFD) framework addresses these challenges by systematically assessing data reliability and relevance to mitigate bias.[60]

*Data commons for data registration.* Data commons provide shared platforms co-locating data, storage, and computing resources with common APIs and tools.[61] They offer centralized environments (e.g., cloud-based tools like Google Colab; https://colab.research.google.com/) where diverse datasets can be managed and analyzed using standardized methods. However, in healthcare, privacy concerns often prevent the open sharing of sensitive data. We envision a data commons model allowing controlled metadata sharing and supporting decentralized analysis, preserving data privacy and autonomy.[62] Standardizing metadata and providing secure tools can enhance data integration efficiency and facilitate robust research findings.

However, CDMs, as seen in initiatives like Sentinel, effectively protect privacy and maintain local data control, important for political and cultural acceptance. The data commons approach must ensure comparable data privacy and governance to be acceptable.

While integrating data commons could benefit the fourth-generation approach by facilitating data registration and discovery, they are not essential. Core objectives—leveraging AI for data access, standardizing data via KGs, and automating analyses—could be achieved without centralized data commons, provided robust interoperability and data governance mechanisms exist.

*From centralized to decentralized study data management.* Initially centralized, the data commons framework has demonstrated significant potential in improving data accessibility and utility across various fields. As highlighted by Guha et al., this framework enables the pooling of vast amounts of public data, making it accessible via standardized APIs.[63] While centralization has its benefits, drug safety is now transitioning toward a decentralized model where data can remain in

its native environment but still be fully accessible and integrated into broader safety surveillance systems. Advancements in Cloud APIs and visualization tools have facilitated this shift, enabling real-time analysis and reporting of safety signals, which could reduce drug-related risks.

Looking ahead, integrating AI—including machine learning for predictive analytics and natural language processing for mining unstructured data (such as social media)—will further enhance the effectiveness of drug safety surveillance. This proactive approach not only addresses current surveillance needs but also helps predict and mitigate risks associated with new pharmaceuticals, ultimately improving patient safety on a global scale.

*Architectural framework for a decentralized data commons.* The proposed architectural framework for a decentralized data commons must support key capabilities, such as data authentication (verifying the publisher and trustworthiness of the data through mechanisms like digital signatures, like those used in open-source software), detailed metadata descriptions (including information about coding scheme versions and the language of the source), and alignment with KGs. This framework should enable data to be described in its native language, and when paired with a large language model (LLM), it allows for the interpretation and integration of diverse data sets without the need for conventional CDM conversions.

*Potential challenges and considerations.* Transitioning to a decentralized data management model introduces several challenges. These include the risk of data loss if a provider withdraws their data from the commons, the potential introduction of low-quality or malicious data by bad actors, and the complexities associated with ensuring continuous data integrity and security. Addressing these challenges requires robust mechanisms for data verification, quality control, and stakeholder cooperation to preserve the integrity and usefulness of the data commons.

*Standardizing data with KGs in drug safety.* The integration of GenAI with KGs represents a significant advancement in addressing key challenges in drug safety, including interoperability, data integration, and predictive analysis. KGs are structured representations that connect entities (e.g., drugs, diseases, clinical trials) through meaningful relationships, transforming complex, multi-dimensional data into actionable insights. By organizing data into interconnected networks, KGs potentially offer a framework for enhancing safety surveillance and enabling predictive, preventive, and personalized medicine. While the current literature on the application of KGs in pharmacovigilance remains sparse, with limited examples demonstrating their utility, this emerging area holds promise for advancing the field and addressing critical gaps in drug safety.

*Integration across diverse datasets.* An important challenge in drug safety is integrating heterogeneous RWD—including clinical, genomic, and chemical datasets—while adhering to FAIR (Findable, Accessible, Interoperable, and Reusable) principles. Currently, many analyses are conducted in silos, requiring significant levels of domain expertise to develop queries that can address specific drug safety questions.

The PheKnowLator project demonstrates the potential of KGs to integrate diverse data modalities, such as genomic, proteomic, clinical, and chemical information, into a unified and FAIR-compliant framework.[64] By building on examples like PheKnowLator, the combination of KGs and GenAI could automate the harmonization of RWD with other high-dimensional datasets. This synergy holds promise for enhancing routine pharmacovigilance activities—such as adverse event detection and monitoring—by enabling faster and more timely access to comprehensive and supporting evidence.

*Predictive models using KGs.* The PlaNet system by Brbić et al. exemplifies the combination of AI and KGs to address safety prediction challenges in clinical trials.[65] PlaNet uses two distinct KGs—one for clinical data and the other for biological and chemical relationships—to predict safety outcomes and adverse events more effectively. By leveraging predictive modeling, PlaNet demonstrates a framework with the potential to improve the accuracy of safety assessments, potentially offering a scalable solution for identifying potential risks earlier in the drug development process.

*Validation and quality assurance.* Despite these advances, ensuring the reliability of AI-driven KG systems remains a critical challenge. Quality assurance processes, such as benchmarking KG models against established safety datasets

and validating predictive outputs, are essential to building trust in their applications. For example, frameworks that compare GenAI predictions to validated adverse event databases can help measure their accuracy, consistency, and reproducibility. Incorporating such validation mechanisms will be crucial as KGs and AI are further integrated into routine safety surveillance systems.

*Toward real-time drug safety surveillance.* The potential for KGs to enable real-time data querying, as demonstrated by LinkedIn's use of KG technologies with sub-millisecond response times,[66] signals a potential, major shift in pharmacovigilance. Applying similar capabilities to drug safety surveillance could allow regulators and researchers to query complex datasets dynamically and efficiently. This is particularly relevant for responding to emerging safety signals in real time, improving decision-making for clinical and regulatory stakeholders.

### *Automatic code generation for PV*

*Innovative applications of GenAI.* GenAI, particularly LLMs, shows promise in transforming healthcare data analysis. LLMs interpret natural language queries and generate coherent responses, bridging the gap between human inquiry and machine-readable data.[67–69] Techniques like retrieval-augmented generation (RAG) further enhance LLMs' accuracy for data-validated responses.[70,71] LLMs can automate complex data interactions, including generating SQL queries for epidemiological research.[72,73]

GenAI can transform federated RWD networks by enabling dynamic evolution of data structures and optimizing queries, reducing the operational burden on data providers. Re-envisioning the extract, transform, and load (ETL) process as a business context document allows for real-time queries executed directly against raw data, bypassing predefined CDM transformations.[74] This approach could enhance the flexibility of data pipelines.

Such advancements allow researchers to ask complex questions across RWD databases, whether in native formats or CDMs. GenAI facilitates comparisons of analyses performed on native versus CDM-transformed data, improving the scope and accuracy of real-world evidence.

For example, assessing a pharmaceutical product's adverse events globally currently requires access to diverse data formats. A GenAI-driven interface could query these sources worldwide, including databases in the United States, EU, Japan, China, and more. Assuming data privacy and governance issues are resolved, researchers could estimate and compare incidence rates across formats and regions, enhancing global result comparability.

These developments suggest a future where GenAI simplifies data handling and broadens analytical capabilities. By offering rapid insights into PV and patient safety, GenAI could extend tools developed for CDM ecosystems across various RWD sources, enhancing overall evidence generation.

*Enhancing data interoperability with AI and KGs.* GenAI and KGs offer promising avenues for data interoperability without relying solely on traditional CDMs. AI-enabled, agent-based data discovery can facilitate real-time analysis across diverse datasets while preserving privacy.[61] KGs standardize data by capturing relationships inherent in heterogeneous sources, reducing information loss and maintaining integrity.[75]

Data commons repositories for data registration can streamline sharing by providing standardized metadata,[62] and automatic code generation tools can enhance code portability and replicability.[76] Brat et al. argue that LLMs' ability to interpret diverse data formats may reduce the reliance on traditional data models, prompting a reevaluation of data interoperability strategies in healthcare.[77]

Collaboration among data custodians, analysts, regulators, and developers is essential to create governance frameworks that balance accessibility with privacy and ethical considerations, moving toward more efficient use of RWD and improving patient outcomes.

Emerging techniques like federated learning and secure multiparty computation provide frameworks for privacy-preserving data analysis, helping ensure sensitive data remains secure while enabling collaborative research.

## *Challenges and future directions*

*Current challenges in CDM implementation.* The growing size and diversity of data networks intensify the challenges of converting data to a standard CDM, compounded by the constantly

changing healthcare landscape.[24] Integrating heterogeneous data sources complicates the creation of universally applicable CDMs essential for drug safety surveillance.[25] Ensuring semantic consistency is crucial to avoid misinterpretations in drug safety analysis.

Kent et al. highlight the difficulties in standardizing data across healthcare systems, where variability in data collection methods can compromise analysis reliability.[26] Overcoming technological, methodological, regulatory, and ethical challenges adds complexity to CDM implementation. Analysts also struggle with accessing data through various vendor tools and employing different strategies, which complicates analyses and prolongs insight generation.[27] In addition, evolving CDMs and their ecosystems can lead to discrepancies in outputs, further complicating comparisons across different models.[28,29]

*Future directions.* Addressing these challenges requires innovative approaches that leverage emerging technologies.

*Enhancing ontological consistency with AI and KGs in biomedical data.* Addressing the complexities of DDNs requires careful consideration of the terminologies and ontologies that underpin the description and categorization of data within these systems. Traditional management of medical terminologies often involves cross-walk maps created by CDMs, but these solutions have inherent limitations, particularly in areas like drug safety, where data must be meticulously accurate, current, and contextually meaningful.

Li et al. introduced FHIR-generative pre-trained transformer (GPT), an innovative approach that combines LLMs with the Fast Healthcare Interoperability Resources (FHIR) standard to enhance health data interoperability.[78] By leveraging LLMs' natural language understanding capabilities, FHIR-GPT can effectively bridge different data formats and terminologies, enabling more seamless data exchange and integration across healthcare systems.[79]

The integration of GenAI and KGs presents new opportunities to overcome these limitations. These technologies offer more dynamic and contextually aware methods for managing the evolution of medical terminologies, ultimately improving the consistency and utility of biomedical data across different systems.

*The role of generative AI in enhancing ontological consistency.* GenAI has the potential to fundamentally improve the management of biomedical terminologies by dynamically harmonizing mismatched terminologies across DDNs. This technology can leverage its ability to analyze vast amounts of structured and unstructured data to identify and resolve inconsistencies in real time.

The reason GenAI holds such promise in this domain is its capacity to generate contextually appropriate mappings that go beyond the rigid structures of traditional crosswalks. Using AI to understand the underlying semantics and relationships between different medical codes, it can provide more accurate and flexible mappings between systems like MedDRA and ICD. This could streamline the integration of data from disparate sources, allowing for faster, more accurate data analysis without the risk of misclassification or oversimplification.

In the observational research landscape, where the quality and reliability of RWD are paramount, GenAI can help standardize the ontological frameworks that underpin these data. This standardization can improve the interoperability of healthcare data systems, ensuring that patient outcomes, safety signals, and treatment efficacy data are consistently represented across multiple platforms and coding standards. Ultimately, this reduces the potential for errors in drug safety surveillance and accelerates the ability to respond to new medical challenges.

GenAI helps to create a dynamic ecosystem where medical terminologies can evolve in sync with the rapid pace of biomedical research, ensuring that healthcare systems are better equipped to handle future pandemics, emerging diseases, and evolving medical knowledge.

*Generative AI and the future of observational research.* The advent of GenAI holds significant potential to revolutionize observational research by enhancing the ETL process and enabling more dynamic data analyses. This technology facilitates direct querying of diverse RWD databases, significantly streamlining the research process by reducing the need for extensive data preparation,

which is crucial for applications like effective signal detection.[80]

GenAI, particularly the use of LLMs, is rapidly advancing in automating and supporting complex computer coding tasks. Recent studies have demonstrated that LLMs when trained with user guides can effectively provide textual responses for coding rules necessary for data entry into databases.[74,81] This suggests a future where LLMs could be trained on specific database structures or data model architectures, enabling them to convert and adapt code in real time for any targeted database system.

These capabilities would dramatically increase the efficiency of database analyses, allowing for optimal coding that minimizes information loss during data conversion processes. This advancement could eliminate the labor-intensive and repetitive tasks associated with data conversion and updates in CDMs or native database structures. Furthermore, this approach would facilitate the effective reuse of code across different DDNs, making it easier to apply proven analytical tools across various data systems without significant reconfiguration.

GenAI tools have the potential to support other aspects necessary for the implementation of observational studies. For example, GenAI can assist with defining, evaluating, and monitoring the consistency of phenotypes over time, as well as summarize information from unstructured observational data for phenotype validation. In addition, the interpretation of study diagnostics to inform study design choices and ultimately choose a design based on a framework that enables causal inference[82] may be supported with GenAI.

Integrating GenAI into observational research will necessitate rigorous quality assurance processes to ensure data integrity and reliability. Leveraging existing frameworks, such as those developed by the OHDSI for quality-assuring CDMs, could provide a foundation for developing AI-specific quality assurance standards. In addition, applying comprehensive quality management systems that are crucial in PV will be essential to build trust and verify the accuracy of AI-generated insights in real-world applications.

By addressing these challenges and ensuring robust validation protocols, GenAI can significantly lower the barriers to efficient and effective observational research, paving the way for more timely insights into patient outcomes and drug safety.

## Conclusion

While CDMs have a current role in record linkage and certain analytical use cases, the advent of AI-driven methods promises a transformative shift in distributed analytics, potentially rendering traditional models less central.

The expansion of electronic healthcare databases and DDNs offers unprecedented opportunities for health data analysis. These tools can significantly improve the identification of drug-related risks and benefits, empowering healthcare professionals to make informed decisions that enhance patient safety. As these systems evolve, they will shape the future of healthcare by leveraging insights from routine clinical data.

Integrating advanced technologies like GenAI and KGs could transform the management and analysis of RWD. Although CDMs are unlikely to be replaced immediately, these technologies indicate a future where CDMs may become less central. KGs can facilitate the analysis of diverse RWD datasets in their native formats, addressing challenges associated with data standardization, privacy, and multi-database analyses.

Tackling ontological challenges from various coding systems is crucial. KGs, supported by GenAI and human-in-the-loop validation, could enable multimodal analyses of RWD alongside other data sources like clinical trials and adverse event reports. This harmonization can lead to more comprehensive and accurate analyses.

GenAI has the potential to revolutionize interactions between data providers and researchers by dynamically managing data structures in federated and decentralized networks. This reduces operational burdens, while KGs provide a robust framework for integrating diverse biomedical data, allowing for deeper insights into drug safety surveillance and personalized medicine.[83,84]

As we adopt these emerging technologies, it is crucial to do so responsibly.[85] Fostering collaboration among data scientists, clinicians, and regulatory bodies, along with ethical AI use, will help maximize the benefits of big data for patient outcomes

and public health. Advanced sandboxing techniques and non-PHI summaries enable analysis without direct patient-level data access, enhancing privacy protections within this new framework.

Despite the long-standing use of CDMs, ongoing discussions in the literature emphasize that no single model offers a perfect solution. A key question is not just how many data sources have been converted into a CDM, but how many remain unavailable and why. Evaluating the value of CDMs and exploring alternative approaches remain central to the discourse, as demonstrated by recent work like that of Tsai et al.[86]

Complexities in accessing multiple data sources—such as the need to use different vendor interfaces for datasets with unique structures—highlight the limitations of current approaches.[27] These challenges could be mitigated if AI-driven solutions allow analysts to focus directly on the data itself, without being constrained by specific data models or interfaces. By enabling a data model-agnostic focus, AI has the potential to simplify and enhance the analytical process across diverse data environments.

In conclusion, the convergence of GenAI and KGs represents a promising frontier in the evolution of drug safety surveillance. By embracing these technologies responsibly and collaboratively, we can address current limitations, unlock deeper insights from complex datasets, and ultimately improve patient outcomes and public health.

## Declarations

### *Ethics approval and consent to participate*

Not applicable, as this is a review solely based on previously published research.

### *Consent for publication*

Not applicable, as this is a review solely based on previously published research.

### *Author contributions*

**Jeffery L. Painter:** Conceptualization; Writing – original draft; Writing – review & editing.

**Darmendra Ramcharran:** Conceptualization; Writing – review & editing.

**Andrew Bate:** Conceptualization; Supervision; Writing – review & editing.

### *Acknowledgements*

None.

### *Funding*

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: GSK covered all costs associated with the conduct of the study, the development of the manuscript, and the decision to publish the manuscript. JP, DR, and AB are employed by GSK and hold financial equities.

### *Competing interests*

The authors declare that there is no conflict of interest.

### *Availability of data and materials*

There are no datasets associated with this perspective.

## ORCID iD

Jeffery L. Painter https://orcid.org/0000-0001-9651-9904

## References

1. Montastruc J-L, Benevent J, Montastruc F, et al. What is pharmacoepidemiology? Definition, methods, interest and clinical applications. *Therapies* 2019; 74(2): 169–174.
2. Jones JK, Van de Carr SW, Rosa F, et al. Medicaid drug-event data: an emerging tool for evaluation of drug risk. *Acta Med Scand* 1984; 215(S683): 127–134.
3. Schneeweiss S and Avorn J. A review of uses of health care utilization databases for epidemiologic research on therapeutics. *J Clin Epidemiol* 2005; 58(4): 323–337.
4. Evans RS. Electronic health records: then, now, and in the future. *Yearb Med Inf* 2016; 25(S01): S48–S61.
5. Klungel OH, Kurz X, De Groot MCH, et al. Multi-centre, multi-database studies with common protocols: lessons learnt from the IMI PROTECT project. *Pharmacoepidemiol Drug Saf* 2016; 25: 156–165.
6. Blettner M, Sauerbrei W, Schlehofer B, et al. Traditional reviews, meta-analyses and pooled analyses in epidemiology. *Int J Epidemiol* 1999; 28(1): 1–9.
7. Stroup DF, Berlin JA, Morton SC, et al. Meta-analysis of observational studies in epidemiology:

a proposal for reporting. *JAMA* 2000; 283(15): 2008–2012.

8. Platt R, Carnahan RM, Brown JS, et al. The US Food and Drug Administration's Mini-Sentinel program: status and direction. *Pharmacoepidemiol Drug Saf* 2012; 21: 1–8.
9. Riera-Guardia N, Saltus CW, Bui CL, et al. *Changes in the landscape of health care database research from 2000 to 2011*. Research Triangle Park, NC: RTI Press. https://doi.org/10.3768/rtipress.2013.RR.0019.1308
10. Chen RT, DeStefano F, Davis RL, et al. The Vaccine Safety Datalink: immunization research in health maintenance organizations in the USA. *Bull World Health Organ* 2000; 78(2): 186–194. https://www.scielosp.org/pdf/bwho/v78n2/v78n2a06.pdf.
11. Chen RT, Glasser JW, Rhodes PH, et al. Vaccine Safety Datalink project: a new tool for improving vaccine safety monitoring in the United States. *Pediatrics* 1997; 99(6): 765–773.
12. Huang C, Jalbert J, Kimura M, et al. The Asian Pharmacoepidemiology Network (AsPEN): promoting multi-national collaboration for pharmacoepidemiologic research in Asia. *Pharmacoepidemiol Drug Saf* 2013; 22: 700–704.
13. Platt RW, Dormuth CR, Chateau D, et al. Observational studies of drug safety in multi-database studies: methodological challenges and opportunities. *eGEMs* 2016; 4(1): 1221.
14. Suissa S, Henry D, Caetano P, et al. CNODES: the Canadian network for observational drug effect studies. *Open Med* 2012; 6(4): e134.
15. Alter GC and Vardigan M. Addressing global data sharing challenges. *J Empirical Res Hum Res Ethics* 2015; 10(3): 317–323.
16. Arlett PR and Kurz X. New approaches to strengthen pharmacovigilance. *Drug Discov Today Technol* 2011; 8(1): e15-e19.
17. Schneeweiss S, Brown JS, Bate A, et al. Choosing among common data models for real-world data analyses fit for making decisions about the effectiveness of medical products. *Clin Pharmacol Ther* 2020; 107(4): 827–833.
18. Yamaguchi M, Inomata S, Harada S, et al. Establishment of the MID-NET medical information database network as a reliable and valuable database for drug safety assessments in Japan. *Pharmacoepidemiol Drug Saf* 2019; 28(10): 1395–1404.
19. Palchuk MB, London JW, Perez-Rey D, et al. A global federated real-world data and analytics platform for research. *JAMIA Open* 2023; 6(2): ooad035.
20. Reisinger SJ, Ryan PB, O'Hara DJ, et al. Development and evaluation of a common data model enabling active drug safety surveillance using disparate healthcare databases. *J Am Med Inform Assoc* 2010; 17(6): 652–662.
21. Toh S, Pratt N, Klungel O, et al. Distributed networks of databases analyzed using common protocols and/or common data models. In: Strom BL, Kimmel SE and Hennessy S (eds) *Pharmacoepidemiology*. Wiley-Blackwell, Chichester, 2019, pp. 617–638.
22. Danese MD, Halperin M, Duryea J, et al. The generalized data model for clinical research. *BMC Med Inform Decis Mak* 2019; 19: 1–3.
23. Roblin DW, Rubenstein KB, Tavel HM, et al. Development of a common data model for a multisite and multiyear study of virtual visit implementation: a case study. *Med Care* 2023; 61: S54–S61.
24. Bourke A, Bate A, Sauer BC, et al. Evidence generation from healthcare databases: recommendations for managing change. *Pharmacoepidemiol Drug Saf* 2016; 25(7): 749–754.
25. Koutkias V. From data silos to standardized, linked, and FAIR data for pharmacovigilance: current advances and challenges with observational healthcare data. *Drug Saf* 2019; 42(5): 583–586.
26. Kent S, Burn E, Dawoud D, et al. Common problems, common data model solutions: evidence generation for health technology assessment. *Pharmacoeconomics* 2021; 39(3): 275–285.
27. Zhou X, Geier J, Shen R, et al. Big data and real world evidence: rapid cycle analysis capability via emerging analytic tools—insights in atopic dermatitis and lessons for wider adoption. *Pharmacoepidemiol Drug Saf* 2019; 28: 68–69.
28. Xu Y, Zhou X, Suehs BT, et al. A comparative assessment of observational medical outcomes partnership and mini-sentinel common data models and analytics: implications for active drug safety surveillance. *Drug Saf* 2015; 38: 749–765.
29. Wang SV, Sreedhara SK and Schneeweiss S. Reproducibility of real-world evidence studies using clinical practice data to inform regulatory and coverage decisions. *Nat Commun* 2022; 13(1): 5126.
30. Cai CX, Halfpenny W, Boland MV, et al. Advancing toward a common data model in ophthalmology: gap analysis of general eye

examination concepts to standard Observational Medical Outcomes Partnership (OMOP) concepts. *Ophthalmol Sci* 2023; 3(4): 100391.

31. Haberson A, Rinner C, Schöberl A, et al. Feasibility of mapping Austrian health claims data to the OMOP common data model. *J Med Syst* 2019; 43: 1–5.
32. Kim H, Choi J, Jang I, et al. Feasibility of representing data from published nursing research using the OMOP common data model. *AMIA Annu Symp Proc* 2016; 2016: 715.
33. Maier C, Lang L, Storf H, et al. Towards implementation of OMOP in a German university hospital consortium. *Appl Clin Inform* 2018; 9(1): 54–61.
34. Zhou X, Murugesan S, Bhullar H, et al. An evaluation of the THIN database in the OMOP Common Data Model for active drug safety surveillance. *Drug Saf* 2013; 36: 119–134.
35. Garza M, Del Fiol G, Tenenbaum J, et al. Evaluating common data models for use with a longitudinal community registry. *J Biomed Inf* 2016; 64: 333–341.
36. Gressler LE, Marinac-Dabic D, Resnic FS, et al. A comprehensive framework for evaluating the value created by real-world evidence for diverse stakeholders: the case for coordinated registry networks. *Therap Innov Regul Sci* 2024; 58: 1042–1052.
37. O'Malley KJ, Cook KF, Price MD, et al. Measuring diagnoses: ICD code accuracy. *Health Serv Res* 2005; 40(5 Pt 2): 1620–1629.
38. Painter JL. Toward automating an inference model on unstructured terminologies: OXMIS case study. *Adv Exp Med Biol* 2010; 680: 645–651.
39. Sawarkar A, Sorbello A, Ripple AML, et al. Detecting adverse drug event safety signals from MEDLINE reports: challenges in employing cross-terminology mapping of MeSH to MedDRA, https://data.lhncbc.nlm.nih.gov/public/mor/pubs/pdf/2017-amia-as-poster.pdf (2017, accessed 2 September 2024).
40. Nadkarni PM and Darer JA. Migrating existing clinical content from ICD-9 to SNOMED. *J Am Med Inform Assoc* 2010; 17(5): 602–607.
41. Brown EG, Wood L and Wood S. The medical dictionary for regulatory activities (MedDRA). *Drug Saf* 1999; 20(2): 109–117.
42. Zhang X, Feng Y, Li F, et al. Evaluating MedDRA-to-ICD terminology mappings. *BMC Med Inform Decis Mak* 2024; 23(Suppl. 4): 299.
43. Khan NF, Harrison SE and Rose PW. Validity of diagnostic coding within the General Practice Research Database: a systematic review. *Br J Gen Pract* 2010; 60(572): e128–e136.
44. Wang Y, Wang L, Rastegar-Mojarad M, et al. Clinical information extraction applications: a literature review. *J Biomed Inform* 2018; 77: 34–49.
45. Livne OE, Schultz ND and Narus SP. Federated querying architecture with clinical & translational health IT application. *J Med Syst* 2011; 35(5): 1211–1224.
46. Lu Z. Information technology in pharmacovigilance: Benefits, challenges, and future directions from industry perspectives. *Drug Healthc Patient Saf* 2009; 1: 35–45.
47. Hall GC, Sauer B, Bourke A, et al. Guidelines for good database selection and use in pharmacoepidemiology research. *Pharmacoepidemiol Drug Saf* 2012; 21(1): 1–10.
48. Cohen JM, Cesta CE, Kjerpeseth L, et al. A common data model for harmonization in the Nordic Pregnancy Drug Safety Studies (NorPreSS). *Norsk Epidemiologi* 2021; 29(1–2). https://doi.org/10.5324/nje.v29i1-2.4053
49. Bate A. Guidance to reinforce the credibility of health care database studies and ensure their appropriate impact. *Pharmacoepidemiol Drug Saf* 2017; 26(9): 1013–1017.
50. McIntosh LD, Juehne A, Vitale CRH, et al. Repeat: a framework to assess empirical reproducibility in biomedical research. *BMC Med Res Methodol* 2017; 17: 1–9.
51. Regulation P. Regulation (EU) 2016/679 of the European Parliament and of the Council. *Off J Eur Union* 2016; 679: 2016.
52. O'Herrin JK, Fost N and Kudsk KA. Health Insurance Portability Accountability Act (HIPAA) regulations: effect on medical record research. *Ann Surg* 2004; 239(6): 772–778.
53. Curtis LH, Weiner MG, Boudreau DM, et al. Design considerations, architecture, and use of the Mini-Sentinel distributed data system. *Pharmacoepidemiol Drug Saf* 2012; 21: 23–31.
54. Kalkman S, Mostert M, Gerlinger C, et al. Responsible data sharing in international health research: a systematic review of principles and norms. *BMC Med Ethics* 2019; 20: 1–13.
55. Walker AM, Zhou X, Ananthakrishnan AN, et al. Computer-assisted expert case definition in electronic health records. *Int J Med Inform* 2016; 86: 62–70.

56. Gini R, Pajouheshnia R, Gardarsdottir H, et al. Describing diversity of real world data sources in pharmacoepidemiologic studies: the DIVERSE scoping review. *Pharmacoepidemiol Drug Saf* 2024; 33(5): e5787.

57. Pajouheshnia R, Gini R, Gutierrez L, et al. Metadata for Data dIscoverability aNd Study rEplicability in obseRVAtional Studies (MINERVA): development and pilot of a metadata list and catalogue in Europe. *Pharmacoepidemiol Drug Saf* 2024; 33(8): e5871.

58. Lavertu A, Vora B, Giacomini KM, et al. A new era in pharmacovigilance: toward real-world data and digital monitoring. *Clin Pharmacol Ther* 2021; 109(5): 1197–1202.

59. Bate A, Chuang-Stein C, Roddam A, et al. Lessons from meta-analyses of randomized clinical trials for analysis of distributed networks of observational databases. *Pharm Stat* 2019 ; 18(1): 65–77.

60. Gatto NM, Campbell UB, Rubinstein E, et al. The structured process to identify fit-for-purpose data: a data feasibility assessment framework. *Clin Pharmacol Therap* 2022; 111(1): 122–134.

61. Grossman RL, Heath A, Murphy M, et al. A case for data commons: toward data science as a service. *Comput Sci Eng* 2016; 18(5): 10–20.

62. Eschenfelder KR and Johnson A. Managing the data commons: controlled sharing of scholarly data. *J Assoc Inf Sci Technol* 2014; 65(9): 1757–1774.

63. Guha RV, Radhakrishnan P, Xu B, et al. Data commons. *arXiv Preprint arXiv:2309.13054*, 2023.

64. Callahan TJ, Tripodi IJ, Stefanski AL, et al. An open source knowledge graph ecosystem for the life sciences. *Sci Data* 2024; 11(1): 363.

65. Brbić M, Yasunaga M, Agarwal P, et al. Predicting drug outcome of population via clinical knowledge graph. *medRxiv*. 2024.

66. Li J, Wade V and Sah M. Developing knowledge models of social media: a case study on LinkedIn. *Open J Semant Web* 2014; 1(2): 1–24.

67. Brown TB. Language models are few-shot learners. *arXiv Preprint arXiv:2005.14165*, 2020, https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf (2020, accessed 2 September 2024).

68. Xie SM, Raghunathan A, Liang P, et al. An explanation of in-context learning as implicit Bayesian inference. *arXiv Preprint arXiv:2111.02080*, 2021.

69. Radford A, Wu J, Child R, et al. Language models are unsupervised multitask learners. *OpenAI Blog*, https://insightcivic.s3.us-east-1.amazonaws.com/language-models.pdf (2019, accessed 2 September 2024).

70. Lewis P, Perez E, Piktus A, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. In: *34th conference on neural information processing systems (NeurIPS 2020)*, Vancouver, BC, Canada, 2020.

71. Topsakal O and Akinci TC. Creating large language model applications utilizing LangChain: a primer on developing LLM apps fast. *Int Conf Appl Eng Nat Sci* 2023; 1: 1050–1056.

72. Zhao P, Zhang H, Yu Q, et al. Retrieval-augmented generation for AI-generated content: a survey. *arXiv Preprint arXiv:2402.19473*, 2024.

73. Ziletti A and D'Ambrosi L. Retrieval augmented text-to-SQL generation for epidemiological question answering using electronic health records. *arXiv Preprint arXiv:2403.09226*, 2024.

74. Painter JL, Chalamalasetti VR, Kassekert R, et al. Automating pharmacovigilance evidence generation: using large language models to produce context-aware structured query language. *JAMIA Open* 2025; 8(1): ooaf003.

75. Hogan A, Blomqvist E, Cochez M, et al. Knowledge graphs. *ACM Comput Surv* 2021; 54(4): 1–37.

76. Lai Y, Li C, Wang Y, et al. DS-1000: a natural and reliable benchmark for data science code generation. In: *International conference on machine learning, PMLR*, Honolulu, Hawaii, 2023, pp. 18319–18345.

77. Brat GA, Mandel JC and McDermott MB. Do We need data standards in the era of large language models? *NEJM AI* 2024; 1(8): AIe2400548.

78. Li Y, Wang H, Yerebakan HZ, et al. FHIR-GPT enhances health interoperability with large language models. *NEJM AI* 2024; 1: AIcs2300301.

79. Namli T, Sınacı AA, Gönül S, et al. A scalable and transparent data pipeline for AI-enabled health data ecosystems. *Front Med* 2024; 11: 1393123.

80. Bate A, Hornbuckle K, Juhaeri J, et al. Hypothesis-free signal detection in healthcare databases: finding its value for pharmacovigilance. *Ther Adv Drug Saf* 2019; 10: 2042098619864744.

81. Painter JL, Mahaux O, Vanini M, et al. Enhancing drug safety documentation search

capabilities with large language models: a user-centric approach. In: *2023 international conference on Computational Science and Computational Intelligence (CSCI)*, Las Vegas, NV, USA, 2023.

82. Hernán MA, Hsu J and Healy B. A second chance to get causal inference right: a classification of data science tasks. *CHANCE* 2019; 32(1): 42–49.
83. Abu-Salih B, Al-Qurishi M, Alweshah M, et al. Healthcare knowledge graph construction: a systematic review of the state-of-the-art, open issues, and opportunities. *J Big Data*. 2023; 10(1): 81.
84. Bean DM, Wu H, Iqbal E, et al. Knowledge graph prediction of unknown adverse drug reactions and validation in electronic health records. *Sci Rep* 2017; 7(1): 16416.
85. Stegmann J-U, Littlebury R, Trengove M, et al. Trustworthy AI for safe medicines. *Nat Rev Drug Discov* 2023; 22(10): 855–856.
86. Tsai DHT, Bell JS, Abtahi S, et al. Cross-regional data initiative for the assessment and development of treatment for neurological and mental disorders. *Clin Epidemiol* 2023; 15: 1241–1252.