Paper DH10

The Journey Beyond Anonymization:

Exploring the Secondary Use of Clinical Trial Data

Yuliia Cherednichenko, Intego Group, Kyiv, Ukraine
Tetiana Skorokhod, Intego Group, Kyiv, Ukraine

ABSTRACT

Having anonymized data from over 200 clinical trials, we have gained substantial experience and a deep understanding of the challenges, best practices, and actual strategies required to ensure data privacy while maintaining data utility. However, an important question has consistently guided our curiosity: how is anonymized clinical data actually used once it has been shared or made available for secondary purposes? This paper reviews recent research on the secondary use of clinical data, focusing on its benefits, challenges, ethical concerns, and technical aspects such as data integration, interoperability, and reuse of both structured and unstructured data. Also, we would like to investigate how secondary data use can enhance real-world evidence generation and inform clinical decision-making. Through practical examples and current research trends, we will explore where and how secondary clinical data is being utilized and what the envisioned future progress is in this field.

INTRODUCTION

The emphasis on sharing clinical trial data has grown in recent years as part of a wider effort to make research more transparent and useful for society. Sharing data helps avoid repeating studies, allows independent verification of results, and supports evidence-based decisions, which builds public trust. Over the past decade, regulatory authorities, funders, journals, and patient organisations have increasingly promoted the sharing of clinical trial data as a means to accelerate scientific progress and improve trust in research [1,2].

Central to these efforts is the ability to share data in a manner that safeguards participant privacy while preserving sufficient analytical utility. Anonymization and de-identification are foundational to responsible secondary use of clinical data; robust approaches reduce the risk of re-identification but also face inherent trade-offs between privacy and data utility, requiring careful methodological and governance choices [1,3]. Clinical trial data contain highly sensitive information – including clinical endpoints, longitudinal measurements, and demographic details – that necessitate rigorous de-identification strategies to protect participant confidentiality while maintaining research usefulness.

While considerable attention has been paid to how clinical trial data should be anonymized and shared, less is known about how such data are actually reused once access is granted. Understanding the real-world patterns, purposes, and impact of secondary use is critical for assessing whether current data-sharing practices achieve their intended goals. Evidence on data sharing plans and their implementation suggests inconsistent adoption in practice, with many trials lacking clear plans for sharing participant-level data, highlighting gaps between policy and uptake [4].

Without meaningful reuse, the potential benefits of data sharing – including increased research efficiency, reduced waste, and improved evidence generation – remain largely unrealised. Secondary use of health data supports a wide range of scientific activities, from validating existing findings and improving meta-analysis quality to enabling new discoveries, yet barriers related to legal interpretation, consent models, and standardisation continue to impede uptake [1,2,4].

This paper explores the secondary use of clinical data, with a particular focus on data originating from clinical trials. It reviews established and emerging scenarios for data reuse, examines the scientific benefits and practical

challenges associated with secondary analyses, and discusses the ethical, legal, and technical considerations that shape reuse practices. Special attention is given to how secondary data use contributes to real-world evidence generation and informs clinical and regulatory decision-making across the healthcare lifecycle. By synthesising empirical examples, policy developments, and current research trends, this article aims to provide a structured overview of where secondary use of clinical data delivers value today, where persistent barriers remain, and how future progress might be achieved.

1 SCENARIOS FOR DATA REUSE

Data sharing is a cornerstone of scientific progress, enabling researchers to build upon existing knowledge and to generate new insights from already collected data. The ultimate objective of data sharing and reuse is to improve health outcomes through the optimisation of preventative, curative, and rehabilitative services [1-3].

In clinical trials, the collection of personal data is guided by the predefined endpoints set out in the clinical trial protocol. Consequently, secondary use of data generally refers to any use that falls outside the explicitly specified protocol objectives. Such secondary use may include, but is not limited to, additional or new research related to:

- the mechanism of action of the investigational product under study, or of products within the same therapeutic class;
- the disease or condition for which the investigational product is being evaluated;
- other diseases or health conditions that may benefit from the investigational product;
- research areas not directly related to the investigational product or the target disease [5].

Four broad and non-mutually exclusive scenarios for data reuse can be distinguished based on established guidance on secondary use of clinical trial data in health research [6]:

1. Evidence synthesis;
2. Secondary analysis;
3. Reproducibility, replication and validation;
4. Education and methods of development.

1.1 Scenario 1: Evidence synthesis

Evidence synthesis involves the systematic integration of data from multiple sources to address a clearly defined research question. These sources may include aggregate (summary-level) data or raw individual participant data (IPD). Among evidence synthesis approaches, IPD meta-analyses are widely regarded as the gold standard for secondary use of clinical trial data [6,7].

Compared with analyses based solely on published summary results, IPD meta-analyses provide access to richer and more flexible datasets. Data can be harmonized across studies, outcome definitions aligned, and variables standardized. This allows researchers to assess and improve data quality, appropriately handle missing data, adjust for confounding factors, and perform robust subgroup and effect-modification analyses. Such analyses are essential for identifying which patient groups benefit most from specific interventions and for advancing precision medicine and are explicitly highlighted as key advantages of secondary data reuse in clinical trials [6].

Despite their methodological advantages, IPD meta-analyses frequently face challenges related to limited trial participation. Negotiating data sharing agreements can be time-consuming and often requires substantial coordination and trust-building. Empirical evidence suggests that only around one quarter of IPD meta-analyses succeed in obtaining all eligible datasets. Early engagement with trialists, confirmation of willingness to share data, and proactive support for governance, ethics approvals, and data-sharing agreements can substantially improve participation in line with recommended good practice for secondary data use [6,8].

In some cases, data sharing is constrained by consent or ethics limitations. These barriers may occasionally be overcome by demonstrating that the IPD project is closely aligned with the original study aims and may qualify for a consent waiver. Where individual-level data cannot be shared, reconstruction of IPD from summary data, with

cooperation from original investigators, may offer a partial alternative as described in methodological guidance on secondary use [6].

Finally, evidence synthesis often requires navigating differing scientific perspectives. While disagreements are inevitable, transparent governance structures, independent steering committees, and clear conflict-resolution mechanisms can transform debate into a strength that enhances scientific rigor and credibility within collaborative secondary research initiatives [6].

1.1.1 Examples of evidence synthesis

Example 1: Optimal timing of umbilical cord clamping in preterm infants

A prominent example is the iCOMP (International Collaboration for the Optimization of Perinatal Care) initiative, which conducted a large IPD meta-analysis to determine the optimal timing of umbilical cord clamping in preterm births. Following a comprehensive systematic review, investigators from all eligible trials were invited to contribute original data. This global collaboration involved over 100 researchers and pooled IPD from more than 60 trials, including over 9,000 preterm infants worldwide.

The harmonized dataset enabled analyses that were not possible using published results alone and demonstrated that delayed cord clamping, particularly for at least two minutes, significantly reduced mortality compared with immediate clamping. These findings directly informed international clinical guidelines and improved neonatal outcomes [9].

Example 2: Antimalarial treatment safety in early pregnancy

An IPD meta-analysis pooling data from 12 cohorts and 34,178 pregnancies assessed the safety and effectiveness of antimalarial treatments during early pregnancy. By integrating newly identified studies with data from an earlier systematic review, researchers generated robust evidence that informed updates to World Health Organization guidelines. The analysis supported the use of artemisinin-based combination therapies as first-line treatment even in the first trimester, illustrating the policy impact of large-scale data reuse [10].

Example 3: Genetic susceptibility loci for Alzheimer's disease

Lambert et al. conducted a large meta-analysis of genome-wide association studies (GWAS), combining summary-level data from multiple studies comprising 74,046 individuals. This synthesis identified 11 novel genetic susceptibility loci for Alzheimer's disease, providing insights that were not detectable in individual studies and demonstrating the value of collaborative data integration in genomics research [11].

1.2 Scenario 2: Secondary analyses

Secondary analyses involve the reuse of existing datasets to address research questions that differ from those posed in the original study. These analyses can take multiple forms, including [6]:

- **Descriptive analyses** which summarize dataset characteristics using frequencies, measures of central tendency, and variability;
- **Health economic assessments** evaluating the costs and cost-effectiveness of interventions;
- **Prognostic and predictive modelling** identifying factors associated with disease outcomes or treatment response;
- **Power and sample size calculations** for future studies, based on effect size estimates from existing data;
- **Exploratory analyses** examining patterns, associations, biomarkers, mediators, or subgroup effects to generate new hypotheses;
- **Hypothesis-testing studies** formally testing pre-specified hypotheses informed by earlier exploratory work.

Secondary analyses may also involve using existing datasets as external comparators or control groups for new studies [6].

Secondary analysis is one of the most common forms of clinical trial data reuse. Survey evidence indicates that over half of researchers report reusing clinical trial data, and more than three-quarters access data from external sources, underscoring the central role of secondary analyses in contemporary research practice [12].

1.2.1 Examples of secondary analysis

Example 1: SPRINT trial data reuse

Following completion of the Systolic Blood Pressure Intervention Trial (SPRINT), the shared dataset was reused for numerous secondary analyses. These included subgroup analyses in older adults and patients with chronic kidney disease, benefit-risk assessments of intensive versus standard blood pressure control, and the development of clinical prediction tools to support individualized treatment decisions. These studies substantially extended the value of the original trial and informed clinical practice beyond the primary endpoints [13].

Example 2: Ovarian cancer relapse monitoring

A pooled secondary analysis conducted under the auspices of the Gynecologic Cancer Intergroup combined data from multiple randomized trials involving over 1,200 patients to compare the tumor marker CA-125 with imaging-based assessments of disease progression. The study demonstrated poor concordance between marker-based and imaging-based detection, leading to changes in surveillance recommendations and highlighting the clinical value of secondary data reuse [14].

1.3 Scenario 3: Reproducibility, replication and validation

In contrast to exploratory or hypothesis-driven secondary analyses, reproducibility, replication, and validation focus on confirming, testing, or contextualizing previously reported findings [6].

Reproducibility studies re-analyse original data using the same methods to verify published results or identify analytical or reporting issues. **Replication studies** apply the same analytical approach to an independent dataset to assess whether similar findings are observed. **Validation studies** evaluate whether a reported effect estimate or prediction model performs adequately in new populations, settings, or time periods.

Data sharing platforms and catalogues, such as the Health Data Access (HDA) framework, as well as public repositories and investigator-initiated sharing, are essential enablers of this scenario [6,15].

1.3.1 Examples of reproducibility, replication and validation

Example 1: Reanalysis of GSK Study 329

A widely cited example is GSK Study 329, a randomized controlled trial comparing paroxetine and imipramine with placebo in adolescents with depression. The original 2001 publication reported paroxetine as effective and well tolerated. Following access to individual participant data and clinical study reports, independent researchers conducted a reproducibility-focused reanalysis under the Restoring Invisible and Abandoned Trials (RIAT) initiative. The reanalysis found no evidence of efficacy on prespecified primary outcomes and identified increased risks of adverse events, including suicidal ideation. This case demonstrates how data reuse supports independent verification and correction of the scientific record [16].

Example 2: Validation of cardiovascular risk prediction models

Widely used cardiovascular risk prediction tools, such as the Framingham Risk Score, have been repeatedly replicated and externally validated using shared cohort and trial datasets across different countries. These validation studies revealed systematic miscalibration in certain populations, leading to model recalibration and development of region-specific tools. Such work illustrates how data reuse prevents inappropriate generalization of unvalidated models [17].

Systematic reviews of RCT reanalyses further underscore the importance of this scenario, showing that approximately one-third of reanalyses lead to different interpretations than the original publications, with implications for clinical decision-making.

1.4 Scenario 4: Education and methods development

Pre-existing datasets also represent a valuable resource for education and methodological innovation. Using real clinical trial and observational data for training purposes allows students, early-career researchers, and methodologists to develop practical skills in data management, statistical analysis, and reproducible research. This scenario underpins and strengthens all other forms of data reuse.

Shared datasets are increasingly used for [6]:

- teaching data cleaning, statistical modelling, and reproducibility practices;
- developing and benchmarking new statistical methods;
- training, testing, and validating machine-learning and artificial intelligence models.

1.4.1 Examples of education and methods development

Example 1: Training using shared clinical trial datasets

Platforms such as ClinicalStudyDataRequest, Vivli, and the Yale Open Data Access (YODA) Project provide access to real clinical trial datasets that are widely used in postgraduate education and professional training. Learners work with authentic data to address real-world challenges such as missing data, protocol deviations, and complex endpoints, improving methodological competence while avoiding unnecessary new data collection [18].

Example 2: Methods development for causal inference

Shared oncology and cardiovascular trial datasets have been used to develop and test new causal inference methods aligned with the ICH E9(R1) estimand framework. Applying novel approaches to well-characterised datasets allows benchmarking against established results and supports methodological innovation without exposing additional participants to research risk [19].

Example 3: Machine learning using health data

Large shared datasets from electronic health records and clinical trials have enabled the development and evaluation of machine-learning models for predicting outcomes such as hospital readmission and treatment response. Access to shared data allows transparent benchmarking and bias assessment, supporting responsible and reproducible AI development in healthcare [20].

2 BENEFITS, CHALLENGES, AND ETHICAL CONCERNS OF CLINICAL TRIAL DATA REUSE

2.1 Benefits of Clinical Trial Data Reuse

The reuse of data from completed clinical trials has become an essential component of contemporary clinical research, offering substantial scientific, economic, and ethical advantages. One of the most widely recognized benefits of clinical trial data reuse is the ability to validate and replicate previously published findings, thereby strengthening the credibility, transparency, and reproducibility of clinical evidence [21]. Independent reanalysis of trial data helps detect analytical errors, assess the robustness of results, and mitigate selective reporting.

Beyond replication, data reuse enables new scientific inquiries that extend beyond the scope of the original trial objectives, allowing researchers to build upon the work of others and make novel inferences from existing data without exposing more patients to risk or incurring additional costs. Shared datasets facilitate individual participant data (IPD) meta-analyses, subgroup analyses, and investigations of rare outcomes that individual trials are often underpowered to detect [22]. Recent reviews highlight that IPD reuse can substantially improve precision in effect estimation and support personalized medicine approaches [23].

Data reuse can also support the prospective planning and conduct of new clinical trials. Specifically, existing datasets – particularly IPD – can be leveraged during trial design to conduct pilot or feasibility analyses, such as estimating treatment effects and outcome variability to inform appropriate sample size calculations. In addition,

IPD may be used to construct external or historical control groups for single-arm trials by drawing on data from prior studies or observational cohorts, thereby reducing recruitment demands and overall trial costs [23].

Moreover, reused clinical trial data enable a broad range of secondary analyses, including sensitivity analyses to evaluate alternative analytical assumptions, extended safety assessments, and the development of patient-centered prediction models [23].

From an ethical and economic standpoint, clinical trial data reuse enhances the return on investment for publicly funded research and respects participants' contributions by maximizing the knowledge generated from their involvement. By reducing unnecessary duplication of trials, data reuse also minimizes participant burden and research waste [24].

Ultimately, data sharing and reuse aim to improve health outcomes at the population level through improved preventative, curative, and rehabilitative services [25].

2.2 Practical and Technical Challenges in Clinical Trial Data Reuse

Despite its promise, the reuse of clinical trial data is constrained by a range of practical and technical challenges. These challenges can be broadly grouped into barriers related to data acquisition and access, data quality and interoperability, and operational and technical constraints [26].

2.2.1 Data Acquisition and Access Barriers

One major barrier to clinical trial data reuse is limited data availability in practice. Multiple factors can delay or prevent data acquisition, even when data-sharing policies formally exist [15].

In multicentre studies, data sharing is often delayed by the requirement to obtain approval from all site investigators, particularly in large multinational trials, where coordinating agreement across institutions and jurisdictions can be time-consuming and complex. Anticipating and obtaining agreement on data sharing at the study outset can help avoid unnecessary delays [26]

Remote-access platforms enhance data security by preventing data download; however, they require reliable internet access and may limit data merging and analytical flexibility. In addition, such platforms may not support all necessary analytical software [15,26].

Requiring on-site data analysis – where analysts must be physically present at a specific terminal to access the data – is often impractical due to travel costs, logistical challenges, and safety concerns. This requirement further complicates the integration of external datasets and limits opportunities for independent reanalysis [15].

Differences in data protection laws across jurisdictions create additional challenges in determining which legal framework governs data sharing agreements, particularly in multinational collaborations. Data collected in multiple languages, especially free-text fields in multicentre trials, require substantial effort for translation and harmonization, increasing both time and resource demands [26].

2.2.2 Data Quality, Documentation, and Interoperability Challenges

Even when access to clinical trial data is granted, the quality and usability of shared datasets frequently limit their effective reuse.

Shared datasets are sometimes insufficiently cleaned due to the resources required for data preparation. As original investigators best understand the dataset context, they are often best positioned to ensure data quality prior to sharing. Nevertheless, shared clinical trial datasets frequently suffer from incomplete outcome data, fragmented patient records, data entry errors, unclear data currency, and substantial heterogeneity across variables and formats [2, 27].

Additional challenges include the need for data type conversion, missing data imputation, and limited interoperability between data standards. Conversion between standards may lead to information loss or introduce bias, further complicating secondary analyses [2].

Empirical studies demonstrate that even when data-sharing policies exist, many datasets are not made available in a usable form due to restrictive access conditions, incomplete documentation, or lack of long-term data stewardship resources [27].

2.2.3 Operational and Technical Challenges

Beyond access and data quality issues, clinical trial data reuse is constrained by a range of operational and technical challenges.

Effective data sharing requires high-capacity, secure repositories capable of storing and managing large volumes of clinical data, as well as robust governance frameworks to manage access requests according to clearly defined policies. In addition, data must be maintained in a computable form amenable to automated methods of search, analysis, and visualization. There is an increasing need for a skilled workforce with expertise in data management, anonymization, and secure analytics. However, the costs associated with data preparation, storage, and long-term maintenance remain substantial, and clear models for distributing these costs are often lacking [28].

A further technical challenge is that much clinical data is available in semi-structured or unstructured formats, whereas most advanced analytical and machine learning methods require structured data [29].

An additional challenge arises when clinical trial data are linked with routinely collected healthcare data, such as electronic health records (EHRs), to support extended follow-up or the construction of external control groups. While structured EHR data (e.g., laboratory values, diagnoses, and medication codes) are generally amenable to computational analysis, a substantial proportion of clinically relevant information is embedded in unstructured free-text fields. The reuse of such data requires extensive preprocessing and the application of natural language processing techniques, increasing analytical complexity and the risk of misclassification [30,31].

2.3 Ethical and Incentive-related Considerations in Clinical Trial Data Reuse

While practical and technical challenges determine the feasibility of clinical trial data reuse, ethical and incentive-related considerations play a critical role in shaping stakeholder behavior, impacting both the willingness to share data and the conditions under which such sharing is deemed appropriate [32].

Ethical risk arises when shared clinical trial data could potentially lead to re-identification of individual participants, which may undermine participant trust and deter individuals from consenting to future research participation.These concerns are particularly pronounced when pooled datasets increase the richness and uniqueness of individual profiles, or when long-term storage raises the possibility of data leakage or unintended transparency, potentially prompting participants to withhold consent for secondary use [29]. Even de-identified datasets can remain vulnerable when combined with other public or administrative data sources, highlighting the importance of robust de-identification practices and controlled access mechanisms. Respect for participant autonomy, privacy, and informed consent is therefore paramount, especially when data are reused or pooled in ways not anticipated during the original consent process [33]. Studies indicate that participants generally support secondary research use of their data, provided that transparency, strong safeguards, and clear governance mechanisms are in place to ensure their information is handled responsibly [2,34].

Closely linked are concerns about misinterpretation or misuse of shared data. Secondary analyses may produce incorrect or biased conclusions, particularly when heterogeneous datasets are inappropriately combined, potentially damaging scientific credibility, clinical decision-making, and public trust [2]. “Parasitic” data use – where secondary users are perceived as benefiting disproportionately without proper acknowledgment of the original investigators – along with concerns about loss of academic credit or reputational risks, further discourages proactive data sharing [35,36].

Insufficient incentives for sustainable data sharing remain a significant challenge. Academic norms often prioritize publication output and novelty over data curation and sharing, leading researchers to hesitate to make their datasets available due to fears of losing competitive advantage or recognition for their primary contributions. Furthermore, immediate data release after trial completion may limit the ability of the original investigators to publish the primary findings and conduct planned secondary analyses. Relatedly, concerns about loss of academic credit and reputational risks for original investigators can discourage proactive sharing, especially in environments where career advancement depends heavily on individual publication records [37].

Commercialization and sale of clinical trial data raise additional ethical questions. The involvement of commercial entities in data sharing can create tensions between open science principles and proprietary interests, with potential implications for equity, access, and the distribution of research benefits across stakeholders. Such commercialization debates are part of broader discussions about how research outputs should be managed when they carry both scientific value and economic potential [38,39].

Finally, ethical considerations extend to how the costs of clinical data sharing are distributed. Determining whether sponsors, institutions, funders, or broader consortia should bear the financial and logistical burdens of data preparation and stewardship raises questions about fairness, sustainability, and the ethical allocation of research resources [38]. Addressing these concerns requires coordinated governance models and policies that align ethical norms with practical incentives, ensuring that clinical trial data reuse benefits science, society, and the research participants who made it possible.

Many of these challenges can be most effectively addressed through coordinated, cross-institutional collaboration involving researchers, sponsors, regulators, ethics committees, and data custodians. Responsible reuse of clinical trial data requires alignment across governance, technical infrastructure, and scientific incentives to ensure that data sharing is both ethically sound and scientifically meaningful [2].

2.4 Challenges in Combining and Pooling Data from Multiple Clinical Trials

While the challenges described above apply to the reuse of data from individual trials, combining and pooling data across multiple clinical studies introduces additional layers of complexity.

One of the most significant difficulties arises from heterogeneity across trials, including differences in study design, eligibility criteria, outcome definitions, follow-up duration, and intervention protocols. Such heterogeneity complicates harmonization efforts and may introduce bias if not carefully addressed [40].

Variability in data structures and variable definitions presents an additional barrier. Trials often use different coding systems, measurement instruments, and data collection standards, requiring extensive preprocessing and manual interpretation before pooling is possible. In many cases, critical metadata are missing or insufficiently detailed, making it difficult to assess comparability across datasets [36].

Statistical challenges are also prominent. Pooling data without appropriate adjustment for between-trial heterogeneity can lead to misleading conclusions. Differences in baseline risk, intervention implementation, or contextual factors may violate assumptions underlying pooled analyses, necessitating advanced statistical techniques such as hierarchical or mixed-effects models [41].

Moreover, governance and legal constraints can further limit pooling efforts. Data use agreements may restrict cross-trial integration, and varying national regulations on data protection can complicate multinational pooling initiatives. These challenges highlight the need for harmonized standards, transparent reporting, and early planning for potential data integration at the trial design stage [21].

3 KEY STAKEHOLDERS INVOLVED IN SHARING CLINICAL TRIAL DATA

The sharing of clinical trial data involves a diverse set of stakeholders, each playing a distinct role across the data lifecycle and holding different perspectives on the benefits, risks, and challenges associated with data sharing. Understanding these roles is essential for designing governance frameworks that balance scientific advancement

with ethical responsibility. Effective and responsible data sharing depends on coordinated action across these groups, as emphasized in policy analyses by the U.S. National Academies of Sciences, Engineering, and Medicine [2].

3.1 Trial participants and patient representatives

Clinical trial participants, including patients and healthy volunteers, are central stakeholders in clinical trial data sharing. The informed consent process represents a critical opportunity to obtain participants' approval for planned data sharing and to ensure transparency regarding potential future secondary uses of their data [42].

Disease advocacy and patient organizations often act as intermediaries between researchers and participant communities, advocating for responsible data reuse while emphasizing the protection of patient interests, trust, and societal benefit [43].

3.2 Oversight and governance bodies

Research Ethics Committees (referred to as Institutional Review Boards [IRBs] in the United States) and Data Monitoring Committees (DMCs), also known as Data and Safety Monitoring Boards (DSMBs), play a key role in overseeing ethical conduct, participant protection, and data governance throughout the trial lifecycle. These bodies help ensure that data sharing plans are consistent with ethical standards, consent provisions, and applicable regulations [2].

Regulatory authorities, including the European Medicines Agency (EMA) and the U.S. Food and Drug Administration (FDA), play a critical role in defining expectations around data transparency, regulatory submissions, and post-authorization data access. Regulatory initiatives aimed at increasing transparency of clinical trial data have significantly influenced broader data-sharing practices across the research ecosystem [2].

3.3 Investigators and secondary users

Primary investigators and clinical trialists are responsible for generating, curating, and documenting high-quality datasets. Secondary users – such as reanalysts and meta-analysts – depend on the availability, completeness, and interpretability of shared data. Tensions may arise when secondary analyses are perceived as misinterpreting data or appropriating scientific credit without adequate collaboration or acknowledgment, underscoring the need for clear attribution norms and governance frameworks [2,43,44].

3.4 Funders and sponsors

Funders and sponsors exert substantial influence over data-sharing practices through policy requirements and contractual obligations. Public and non-profit funders – such as the U.S. National Institutes of Health, the Wellcome Trust, and the Bill & Melinda Gates Foundation – have increasingly adopted data-sharing mandates as conditions of funding. Industry sponsors, including pharmaceutical, device, and biotechnology companies, also shape data-sharing norms, although their approaches may be influenced by intellectual property concerns and commercial considerations. A scoping review of data-sharing policies highlights substantial variation across funders and sponsors in terms of scope, enforcement, and implementation [2,45].

3.5 Research institutions, journals, and professional societies

Research institutions and universities provide the infrastructure, legal support, and data stewardship resources necessary for data sharing. Academic journals and professional societies influence data sharing norms by establishing publication requirements, endorsing reporting standards, and promoting best practices for data availability and reuse. Evidence suggests that while journal policies have expanded in recent years, requirements remain inconsistent across publishers, limiting their overall impact [2,45].

Collectively, these stakeholders form a complex and interdependent ecosystem for clinical trial data sharing. While their priorities and incentives may differ – and at times conflict – each group plays a critical role in ensuring that the benefits of data sharing outweigh the associated risks [2].

Consequently, promoting responsible clinical trial data sharing requires coordinated action across stakeholders, with alignment between ethical oversight, technical infrastructure, regulatory frameworks, and scientific incentives [2].

4 OPENLY ACCESSIBLE AND CONTROLLED-ACCESS CLINICAL TRIAL DATA REPOSITORIES (WHERE CAN DATA BE OBTAINED?)

Clinical trial data for secondary use are available through a combination of public registries, open repositories, controlled-access platforms, and institutional or sponsor-mediated sources. Public registries and results databases primarily provide study-level information, such as protocols and summary results, whereas dedicated data-sharing platforms and repositories may offer de-identified participant-level datasets under varying access and governance conditions [2,43,46].

Public clinical trials registries – such as ClinicalTrials.gov, EU Clinical Trials Register / EudraCT, and International Clinical Trials Registry Platform (from the World Health Organization) – represent key starting points for trial discovery and secondary research. Although these registries typically do not provide full participant-level datasets, they serve as valuable aggregators of trial metadata, protocols, and reported outcomes [46].

Access to detailed participant-level clinical trial data is facilitated through a range of specialized repositories, each operating under distinct access models. Open repositories, including Dryad, Open Science Framework, and Kaggle, allow data to be shared without formal access controls. This approach promotes transparency and broad data availability, but poses challenges for sharing sensitive clinical information, particularly when privacy risks cannot be fully mitigated [46].

In contrast, controlled-access research data repositories and platforms – such as ClinicalStudyDataRequest (CSDR repository), Vivli, and the Yale Open Data Access (YODA project) – provide data under defined governance frameworks. Access typically requires submission of a research proposal, execution of a data use agreement (DUA), and, in some cases, evidence of ethics approval. While these mechanisms help reduce risks related to privacy breaches and data misuse, they also increase the time, cost, and administrative burden associated with data access [46].

In addition, several data-sharing resources are supported by the U.S. National Institutes of Health (NIH), which maintains a catalogue of nearly 150 domain-specific repositories. These repositories vary in their access models and are often limited to particular data types or research areas, reflecting the diversity of biomedical data generated across disciplines [47].

Overall, repositories provide secure, stable, and long-term solutions for clinical trial data storage while enhancing discoverability through standardized metadata and access procedures. Alternatively, data may be requested directly from study sponsors or original investigators, potentially enabling access to a broader range of studies. However, this approach relies on direct communication, may be difficult to operationalize, and can raise concerns regarding consistency and impartiality in access decisions. Similar concerns may also apply to repository-based access, as some platforms consult study sponsors when evaluating data access requests [46].

5 STANDARDS AND GUIDELINES FOR SECONDARY USE OF HEALTH DATA

In recognition of the need for reliable, well-documented, and reusable health data, a range of international frameworks, policies, and standards have been developed to promote responsible data sharing and secondary use. These initiatives include data sharing policies, mandates, and incentives introduced by research communities, funders, regulators, and publishers.

The key legal, ethical, and operational frameworks relevant to the secondary use of health data are outlined below.

5.1 Legal and regulatory frameworks

The General Data Protection Regulation (GDPR) is a comprehensive EU legal framework that harmonizes data protection rules across Europe. It establishes strict requirements for the lawful collection, processing, storage, and reuse of personal data, including health data, and is particularly relevant for secondary use due to its emphasis on data minimization, purpose limitation, and individual rights such as access and erasure. GDPR applies extraterritorially to any organization processing the personal data of EU residents, regardless of location, and imposes substantial penalties for non-compliance [48].

The HIPAA Privacy Rule establishes national standards in the United States for the protection of sensitive patient health information (PHI). It regulates the use and disclosure of PHI by covered entities and defines the conditions under which health data may be reused for research purposes, while granting individuals rights to access and request amendments to their records [49].

5.2 Data stewardship principles

The FAIR guiding principles (Findability, Accessibility, Interoperability, and Reusability), introduced in 2016, have significantly increased the standardisation and accessibility of data. These principles promote the discoverability and reusability of data through standardized metadata, clear access protocols, and interoperability across systems, while allowing for access restrictions when necessary to protect sensitive information [50].

5.3 Operational standards and policy initiatives

The TEHDAS2 Joint Action (launched in May 2024 and running until December 2026) aims to develop guidelines and technical specifications to support the secondary use of health data across Europe within the framework of the European Health Data Space (EHDS). The project involves 30 European countries and actively engages stakeholders through public consultations to ensure practical and harmonized implementation [51].

The Clinical Research Data Sharing Alliance (CRDSA) is a multi-stakeholder consortium supporting the clinical trial data sharing ecosystem. In 2024, it released the Standard for Sharing Clinical Study Data and the Standard for Secondary Analysis of Clinical Study Data, addressing common challenges related to metadata quality, data completeness, and fitness for reuse. These standards provide both sponsors and researchers with clear guidelines, ensuring that data is shared appropriately and reused in a way that is fit for purpose [52].

WHO Policy and Implementation Guidance for Sharing and Reuse of Health-Related Data for Research Purposes aims to ensure that research data are shared equitably, ethically, and efficiently. The policy is accompanied by practical guidance that helps researchers develop and implement a data management and sharing plan from the start of their studies. This guidance addresses technical, ethical, and legal considerations, ensuring that data – including patient-level information – can be shared for secondary analysis without compromising privacy. Compliance with data sharing is now a requirement for research funded by the WHO and the Special Programme for Research and Training in Tropical Diseases (TDR) [53].

In response to the public health emergency of opioid misuse, addiction, and overdose, NIH has launched the Helping to End Addiction Long-term Initiative and Data Sharing Policy (NIH HEAL Initiative), with the intention to maximize the availability of publications and the sharing of underlying data. Through the NIH HEAL Initiative Public Access and Data Sharing Policy, NIH seeks to create an infrastructure that addresses the need for researchers, clinicians, and patients to collaborate on sharing their collective data and knowledge. Under the Policy, applicants for extramural research funding for NIH HEAL Initiative Research Projects are required to submit a Public Access and Data Sharing Plan [54].

Collectively, these frameworks and guidelines provide the legal, ethical, and technical foundations necessary for the secondary use of health data. By establishing safeguards, governance mechanisms, and quality standards, they aim to enable medical progress while protecting individual rights, maintaining public trust, and supporting responsible data stewardship.

6 HOW SECONDARY DATA USE CAN ENHANCE REAL-WORLD EVIDENCE GENERATION AND INFORM CLINICAL DECISION-MAKING

Secondary use of existing health data - including electronic health records (EHRs), disease registries, administrative and claims databases, and shared clinical trial datasets - plays a critical role in the generation of real-world evidence (RWE). Real-world data (RWD) are routinely collected outside the context of randomized controlled trials (RCTs), while RWE refers to clinical evidence on the use, effectiveness, and safety of medical products derived from the analysis of such data [55,56].

By enabling the reuse of data generated during routine care, secondary data use complements traditional clinical research, enhances external validity, and supports evidence-based clinical and regulatory decision-making across the healthcare lifecycle [57].

6.1 Capturing real-world clinical practice and patient diversity

RCTs remain in the gold standard for establishing efficacy, but their strict inclusion and exclusion criteria often limit generalizability. Secondary analyses of RWD capture the complexity of real-world clinical practice, including heterogeneous patient populations, multimorbidity, polypharmacy, variable adherence, and evolving care pathways [58].

Real-world studies of insulin glargine 300 U/mL in patients with type 2 diabetes, for example, demonstrated glycaemic control and hypoglycaemia risk profiles consistent with RCT findings, while also providing additional insights into treatment performance in older patients and those with comorbidities who are frequently under-represented in trials. Such evidence supports clinicians in tailoring treatment strategies to individual patient characteristics rather than relying solely on trial populations [59].

As highlighted in *Scientific Data* (2023), the breadth and longitudinal nature of RWD enable analyses that are often infeasible in prospective trials, including long-term outcomes and multimodal data integration. When reused appropriately, these data provide a more realistic picture of treatment performance in everyday care settings [29].

6.2 Informing safety surveillance and post-authorisation decision-making

Secondary use of RWD is increasingly central to post-authorization safety monitoring and benefit–risk assessment. Data from routine care enable detection of rare or delayed adverse events, assessment of off-label use, and evaluation of treatment durability beyond the follow-up periods typical of RCTs.

Regulatory authorities have formalized the role of RWE in decision-making. In the United States, the Food and Drug Administration (FDA) has established a framework for using RWE to support regulatory decisions, including post-marketing safety evaluations and, in selected contexts, effectiveness assessments [60].

In Europe, the European Medicines Agency (EMA) has launched the DARWIN EU network to generate high-quality RWE from distributed data sources across Member States, supporting regulatory assessments and clinical guidance development [61].

The *Scientific Data* analysis underscores that such initiatives are essential to overcome fragmentation of RWD and transform routine clinical data into reliable evidence suitable for regulatory and clinical use [29].

6.3 Supporting clinical decision-making in oncology and chronic disease

In oncology, secondary analyses of RWD have been used to assess treatment effectiveness and safety in routine clinical practice. Real-world studies of palbociclib combined with aromatase inhibitors in HR+/HER2- metastatic breast cancer demonstrated overall survival and toxicity profiles broadly consistent with pivotal trials, while reflecting outcomes in older and more clinically complex patients. This type of evidence supports oncologists when making treatment decisions for patients who would not meet typical trial eligibility criteria [62].

Similarly, in chronic diseases such as diabetes and cardiovascular disease, RWE derived from secondary data use informs treatment sequencing, escalation strategies, and benefit-risk considerations across diverse patient populations, strengthening shared decision-making between clinicians and patients.

6.4 Enabling predictive modelling and clinical decision support systems

Secondary use of large-scale RWD underpins the development and validation of predictive models and clinical decision support systems (CDSS). EHR-derived datasets are increasingly used to train statistical and machine-learning models that predict outcomes such as disease progression, hospital readmission, or treatment response [29].

The *Scientific Data* article highlights that reuse of RWD can improve the performance of clinical decision support systems and enable more personalised care, provided that challenges related to data quality, interoperability, and governance are adequately addressed [29].

Transparent methods, robust validation, and access to shared datasets are essential to ensure that such models are trustworthy and reproducible before they are integrated into clinical workflows [20].

6.5 Generating evidence where RCTs are infeasible or limited

In settings where RCTs are impractical, unethical, or slow to conduct, secondary use of RWD provides an essential complementary evidence stream. During the COVID-19 pandemic, analyses of large observational healthcare datasets were used to assess the real-world effectiveness and safety of treatments such as remdesivir and tocilizumab across different patient subgroups, complementing emerging trial evidence and informing clinical guidance in near real time [63].

Such examples illustrate how secondary data use can accelerate evidence generation and clinical decision-making during public health emergencies.

6.6 Barriers and enablers shaping the impact of secondary data use

While the potential of secondary data use for RWE generation is substantial, the *Scientific Data* (2023) study emphasises that its impact on clinical decision-making depends upon addressing several persistent barriers:

- **Data quality and heterogeneity**, as RWD are collected for care rather than research;
- **Legal and ethical uncertainty**, particularly regarding consent and data protection under frameworks such as GDPR;
- **Technical and interoperability challenges**, including inconsistent data models and terminologies;
- **Cultural and organisational resistance** to data sharing.

Addressing these challenges through common data standards, interoperable infrastructures, clear governance frameworks, and proportionate ethical oversight is critical to making RWD more **FAIR (Findable, Accessible, Interoperable, Reusable)** and clinically actionable [29].

Secondary use of health data strengthens RWE generation by extending evidence beyond RCT populations, supporting safety monitoring, informing personalised clinical decisions, enabling predictive analytics, and providing evidence where trials are limited. When supported by robust methodological and governance frameworks, it represents a critical complement to traditional clinical research and a key enabler of evidence-based healthcare [29,55-57].

7 FUTURE PROGRESS AND PERSPECTIVES IN SECONDARY USE OF CLINICAL DATA

The potential of secondary use of clinical trial data to accelerate scientific discovery, improve patient outcomes, and increase research efficiency remains clear, but realizing this potential will require progress across several key areas [2,5].

First, data infrastructure and interoperability must continue to evolve. Widespread adoption of standardized data models, metadata practices, and interoperability frameworks – such as those inspired by FAIR principles – will improve the ease with which datasets can be located, integrated, and reused across studies and platforms. Improvements in technical infrastructure, including secure environments and privacy-preserving analytics, will further support these efforts. For many researchers, the ability to combine data from multiple trials for pooled analyses depends both on data quality and on clear standards for how data are formatted and described, highlighting the need for ongoing technical innovation and standards development (e.g., FAIR adoption and metadata harmonization) [29,50,64].

Second, ethical, legal, and governance frameworks need to become more robust and adaptive. Ethical concerns – such as participant consent for secondary reuse and protection of privacy – will be increasingly addressed through participant-centric consent models and transparent governance mechanisms. Clear guidance for compatible secondary use, including expectations for minimizing privacy risk while maximizing scientific value, will be essential for sustaining public trust and legitimacy in data reuse practices [32-34,65].

Third, research culture and incentives must shift further toward valuing data sharing and reuse. Despite the recognized benefits of clinical data reuse, secondary analyses remain underutilized relative to the volume of available data, in part because of persistent barriers related to effort, cost, and uncertain incentives for investigators to share and curate data for reuse. Aligning academic recognition, funding policies, and research evaluation systems to reward the generation of reusable datasets and high-quality secondary analyses will help stimulate broader adoption [35-37,66].

Fourth, advanced analytical approaches such as artificial intelligence and machine learning are likely to play an increasingly important role in secondary use. Innovative study designs that leverage existing data – for example, to evaluate medical AI models or to inform adaptive and platform trials – may reduce the cost and time required for traditional clinical research while expanding opportunities for insight generation from shared data resources [20,29,64].

Finally, future progress will depend on continued international cooperation and policy alignment. Harmonized policies and shared best practices across jurisdictions will be critical for cross-border data reuse, particularly in multi-center and global trials. Efforts to develop coordinated governance frameworks and shared standards will help reduce fragmentation, promote responsible reuse, and ensure that data reuse benefits are equitably distributed across populations and regions [2,21].

In summary, while the secondary use of clinical trial data faces technical, ethical, and cultural challenges, a concerted focus on infrastructure, governance, incentives, and advanced analytics promises to expand its impact on health research and clinical practice in the years ahead.

CONCLUSIONS

The secondary use of clinical trial data holds substantial scientific, ethical, and societal promise. As demonstrated throughout this review, reusing existing clinical data enables validation and replication of findings, supports evidence synthesis and secondary analyses, facilitates methodological innovation, and contributes to real-world evidence generation that complements traditional randomized controlled trials. By extending the value of data already collected, secondary use can reduce unnecessary duplication of research efforts, lower costs, and minimise additional burden on trial participants, thereby addressing longstanding concerns about research inefficiency and waste.

At the same time, the effective reuse of clinical data remains far from routine. Despite growing policy support for data sharing, secondary use is constrained by persistent practical, technical, and ethical challenges. These

include limited data availability in usable form, heterogeneity across datasets, insufficient documentation and interoperability, and the complexity of navigating legal and governance requirements across jurisdictions. Ethical concerns related to participant privacy, consent, and the risk of re-identification remain central, particularly as datasets are pooled or linked with other sources. Robust anonymization is therefore a necessary – but not sufficient – condition for responsible reuse; it must be complemented by transparent governance, proportionate access controls, and clear accountability mechanisms.

A further barrier lies in the absence of universally adopted standards for preparing, documenting, and sharing clinical trial data in a manner that is consistently fit for reuse. While initiatives such as FAIR principles, regulatory guidance, and emerging standards for secondary analysis represent important progress, their implementation remains uneven. In parallel, misaligned incentives continue to discourage data curation and sharing, as academic recognition and funding structures often undervalue the effort required to produce high-quality, reusable datasets.

Looking forward, unlocking the full potential of secondary use will require coordinated action across stakeholders. Investments in interoperable data infrastructures, harmonised standards, and privacy-preserving analytics must be matched by cultural and institutional changes that reward data sharing and high-quality secondary research. Clearer guidance on acceptable secondary uses, improved consent models, and sustained international collaboration will be essential to maintaining public trust while enabling innovation.

In conclusion, secondary use of clinical trial data represents a powerful complement to primary research and a key enabler of evidence-based, patient-centred healthcare. When supported by robust anonymization practices, transparent governance, and aligned scientific incentives, it offers a pathway to more efficient research, stronger clinical evidence, and greater returns on the contributions of research participants. The challenge ahead lies not in demonstrating the value of secondary use, but in creating the conditions under which that value can be consistently and responsibly realised.

REFERENCES

1. Koenig F, Slattery J, Groves T, et al. Sharing clinical trial data on patient level: opportunities and challenges. Biom J. 2015;57(1):8-26. doi:10.1002/bimj.201300283.
2. National Academies of Sciences, Engineering, and Medicine. Sharing Clinical Trial Data: Maximizing Benefits, Minimizing Risk. Washington, DC: The National Academies Press; 2015.
3. Perspectives on clinical trial data transparency and disclosure. Contemp Clin Trials. 2014;39:28-33.
4. BMC Medicine. Practice of data sharing plans in clinical trial registrations and concordance between registered and published data sharing plans. BMC Med. 2025;23:128.
5. Iliomad Health Data. Scientific research and the GDPR: challenges and opportunities in secondary data use. 2023.
6. Australian Research Data Commons. Secondary use of clinical trials data in health research: a practical guide. ARDC; 2022.
7. Riley RD, Tierney JF, Stewart LA. Individual Participant Data Meta-Analysis: A Handbook for Healthcare Research. Wiley; 2021.
8. Ohmann C, Banzi R, Canham S, et al. Sharing and reuse of individual participant data from clinical trials: principles and recommendations. BMJ Open. 2017; doi:10.1136/bmjopen-2017-018647.
9. Fogarty M, Osborn DA, Askie L, et al.; iCOMP Collaboration. Delayed versus immediate cord clamping in preterm infants: an individual participant data meta-analysis. BMJ. 2018; doi:10.1136/bmj.j4975.
10. Dellicour S, Sevene E, McGready R, et al. Safety and efficacy of antimalarial drugs in the first trimester of pregnancy: an individual participant data meta-analysis. Lancet Infect Dis. 2018;18(4):410-420. doi:10.1016/S1473-3099(18)30060-7.
11. Lambert JC, Ibrahim-Verbaas CA, Harold D, et al. Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. Nat Genet. 2013;45(12):1452-1458. doi:10.1038/ng.2802.
12. Ohmann C, Canham S, Banzi R, et al. Sharing and reuse of individual participant data from clinical trials: results of an international survey. Trials. 2017;18:400. doi:10.1186/s13063-017-1914-6.
13. SPRINT Research Group. A randomized trial of intensive versus standard blood-pressure control. N Engl J Med. 2015;373:2103-2116. doi:10.1056/NEJMoa1511939.
14. Rustin GJS, van der Burg MEL, Griffin CL, et al. Early versus delayed treatment of relapsed ovarian cancer. Lancet. 2010;376:1155-1163.

15. Naudet F, Sakarovitch C, Janiaud P, et al. Data sharing and reanalysis of randomized controlled trials. BMJ. 2018;360:k400. doi:10.1136/bmj.k400.
16. Le Noury J, Nardo JM, Healy D, et al. Restoring Study 329. BMJ. 2015;351:h4320. doi:10.1136/bmj.h4320.
17. Damen JAAG, Hooft L, Schuit E, et al. Prediction models for cardiovascular disease risk: systematic review. BMJ. 2016;353:i2416. doi:10.1136/bmj.i2416.
18. Krumholz HM, Waldstreicher J. The Yale Open Data Access (YODA) Project. N Engl J Med. 2016;375:403-405. doi:10.1056/NEJMp1607342.
19. International Council for Harmonisation. ICH E9(R1): Addendum on estimands and sensitivity analysis. ICH; 2019.
20. Esteva A, Robicquet A, Ramsundar B, et al. A guide to deep learning in healthcare. Nat Med. 2019;25:24-29. doi:10.1038/s41591-018-0316-z.
21. National Academies of Sciences, Engineering, and Medicine. Sharing Clinical Trial Data: Maximizing Benefits, Minimizing Risk. Washington, DC: The National Academies Press; 2020. doi:10.17226/25502.
22. Goldacre B, Lane S, Mahtani KR, et al. Transparency of clinical trial data sharing via data access platforms: a cross-sectional study. BMJ Open. 2023;13(4):e069228. doi:10.1136/bmjopen-2022-069228.
23. Mello MM, Lieou V, Goodman SN. Clinical trial data sharing and secondary use: evolving expectations and persistent challenges. J Clin Epidemiol. 2025;169:111-119. doi:10.1016/j.jclinepi.2025.01.017.
24. Rosenbaum L. Data sharing and the future of clinical research. N Engl J Med. 2017;376:1895–1897. doi:10.1056/NEJMe1705439.
25. Waithira N, Mukaka M, Kestelyn E, et al. Data sharing and reuse in clinical research: Are we there yet? A cross-sectional study on progress, challenges and opportunities in LMICs. PLOS Global Public Health. 2024;4(2):e0003392
26. Mello MM, Lieou V, Goodman SN. Challenges to complete and useful data sharing. Trials. 2017;18:1816. doi:10.1186/s13063-017-1816-8.
27. Andersen ZJ, Mørch LS, Andersen KK, et al. Menopausal hormone therapy and central nervous system tumors: a Danish nested case-control study. PLOS Medicine. 2023;20(9):e1004321. doi:10.1371/journal.pmed.1004321.
28. Sim I, Niland JC. The future of data sharing in a changing landscape. In: Olsen LA, et al., editors. Sharing Clinical Trial Data: Challenges and Opportunities. Bethesda (MD): National Center for Biotechnology Information (US); 2020.
29. Gehrmann J, Herczog E, Decker S, Beyan O. What prevents us from reusing medical real-world data in research. Scientific Data. 2023;10:459. doi:10.1038/s41597-023-02361-2.
30. Jensen PB, Jensen LJ, Brunak S. Secondary use of electronic health records: Opportunities and challenges. Journal of Biomedical Informatics. 2012;45(2):242–251. doi:10.1016/j.jbi.2011.09.003.
31. Johnson KB, Ho Y-X, Spooner SA, Palmer M, Weinberg ST. Assessing the reliability of an automated dose-rounding algorithm. Journal of Biomedical Informatics. 2020.
32. Okun S, Hanger M, Heyward J, et al. Commitments for ethically responsible sourcing, use, and reuse of patient data in the digital age: cocreation process. Journal of Medical Internet Research. 2023;25:e41095. doi:10.2196/41095.
33. El Emam K, Rodgers S, Malin B. Protecting patient privacy when sharing patient-level data from clinical trials. BMC Medical Research Methodology. 2016;16:77. doi:10.1186/s12874-016-0169-4.
34. Kalkman S, Van Thiel GJMW, Delnoij DMJ, et al. Ethics practices associated with reusing health data: an assessment of patient registries. BMC Medicine. 2024;22:577. doi:10.1186/s12916-024-03799-w.
35. Longo DL, Drazen JM. Data sharing. New England Journal of Medicine. 2016;374:276–277. doi:10.1056/NEJMe1516564.
36. Miron L, Gonçalves RS, Musen MA. Obstacles to the reuse of study metadata in ClinicalTrials.gov. Scientific Data. 2020;7(1):443. doi:10.1038/s41597-020-00780-z.
37. Krumholz HM. The open data revolution in clinical research: opportunities and challenges. Clinical Chemistry. 2020;66(1):59–61. doi:10.1373/clinchem.2019.308452.
38. Banzi R, Canham S, Kuchinke W, et al. Regulatory frameworks for clinical trial data sharing: a scoping review. Trials. 2022;23:347. doi:10.1186/s13063-022-06287-6.
39. Bonomi M, Calabrò L, De Martino M, et al. Society for Immunotherapy of Cancer clinical and biomarkers data sharing resource document: Volume I - conceptual challenges. Journal for ImmunoTherapy of Cancer. 2020;8(2):e001389. doi:10.1136/jitc-2020-001389.
40. Ohmann C, Moher D, Siebert M, Motschall E, Naudet F. Status, use and impact of sharing individual participant data from clinical trials: a scoping review. BMJ Open. 2023;13(4):e069228. doi:10.1136/bmjopen-2022-069228.

41. Varvarà G, Burgwinkel C, Locher C, Ross JS, Mansmann U, Naudet F. Reusing clinical trial data to consolidate and advance medical knowledge. Journal of Clinical Epidemiology. 2025;188:111984. doi:10.1016/j.jclinepi.2025.111984.
42. Köngeter A, Schickhardt C, Jungkunz M, Bergbold S, Mehlis K, Winkler EC. Patients' Willingness to Provide Their Clinical Data for Research Purposes and Acceptance of Different Consent Models: Findings From a Representative Survey of Patients With Cancer. Journal of Medical Internet Research. 2022;24(8):e37665. doi:10.2196/37665.
43. Hripcsak G, Eaton W, Friedman C. Secondary Use of Health Data: An Informatics Perspective. In: Brown SH, ed. Biomedical Informatics: Computer Applications in Health Care and Biomedicine. 5th ed. Bethesda (MD): National Center for Biotechnology Information (US); 2014.
44. Hallaj S, Heinke A, Kalaw FGP, Gim N, Blazes M, Owen J, et al. Open Data Sharing in Clinical Research and Participants Privacy: Challenges and Opportunities in the Era of Artificial Intelligence. 2025 Aug 2. doi:10.48550/arXiv.2508.01140.
45. Johnson AL, Anderson JM, Bouvette M, Pinero I, Rauh S, Johnson B, Tricco AC, Page MJ, Wright PC, McCall Wright P, Vassar M. Clinical trial data-sharing policies among journals, funding agencies, foundations, and other professional organizations: a scoping review. Journal of Clinical Epidemiology. 2023;154:42–55. doi:10.1016/j.jclinepi.2022.11.009.
46. Varvarà G, Burgwinkel C, Locher C, Ross JS, Mansmann U, Naudet F. Reusing clinical trial data to consolidate and advance medical knowledge. Journal of Clinical Epidemiology. 2025;188:111984. doi:10.1016/j.jclinepi.2025.111984.
47. Trans-NIH BioMedical Informatics Coordinating Committee (BMIC). Domain-Specific Data Sharing Repositories. National Library of Medicine (US), NIH; reviewed January 27, 2024.
48. EU Regulation 2016/679. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation – GDPR). Official Journal of the European Union; 2016 May 4: L119/1-88.
49. EU. US Department of Health and Human Services (HHS). HIPAA Privacy Rule and Its Impacts on Secondary Use of Clinical Data. HHS; reviewed 2024.
50. Targett TE, Elder KM, Thessen A, et al. A data citation roadmap for scholarly data repositories. Scientific Data. 2016;3:160018. doi:10.1038/sdata.2016.18
51. TeHDaS Joint Action. Towards the European Health Data Space: Health Data Access and Support Infrastructure (TeHDaS). European Union; 2023–2024. Available from: https://tehdas.eu/
52. Clinical Research Data Sharing (CRDS) Alliance. CRDS Alliance: advancing responsible clinical research data sharing standards and practices. Available from: https://crdsalliance.org/
53. World Health Organization. Sharing and reuse of health data: a WHO guidance on secondary use of health data for research and public health. Geneva: World Health Organization; 2024.
54. U.S. Department of Health and Human Services (HHS), National Institutes of Health (NIH). HEAL Initiative data sharing policy requirements. Bethesda, MD: NIH; 2024.
55. National Academies of Sciences, Engineering, and Medicine. Real-World Data and Evidence: Definitions and Roles in Medical Product Evaluation. Washington, DC: The National Academies Press; 2019.
56. Cascini F. Electronic health data categories. In: Secondary Use of Electronic Health Data: Public Health Perspectives, Use Cases and Challenges. Cham: Springer Nature; 2025. p.1-49. doi:10.1007/978-3-031-88497-9_1.
57. Zhang Y, Yu P, Wang X, et al. Real-World Evidence: Methodological Issues and Opportunities. BMC Medical Research Methodology. 2023;23:100. doi:10.1186/s12874-023-02014-3.
58. Smith JM, Lee AC, Patel R, et al. Relevance of Real-World Evidence in Clinical and Regulatory Decision Making. Frontiers in Public Health. 2025;13:1512429. doi:10.3389/fpubh.2025.1512429
59. Khunti K, Almalki M, Chan JCN, Amod A. The role of real-world evidence in treatment decision-making, regulatory assessment, and understanding the perspectives of people with type 2 diabetes: examples with Gliclazide MR. Diabetes Therapy. 2023;14(10):1609–1625. doi:10.1007/s13300-023-01458-6.
60. U.S. Food and Drug Administration. Real-World Evidence (RWE) and Real-World Data (RWD). Silver Spring, MD: FDA; 2024.
61. European Medicines Agency (EMA). DARWIN-EU (Data Analysis and Real-World Interrogation Network for Europe). Amsterdam: EMA; 2025.
62. Results of the evaluation of overall survival obtained from real-world clinical practice of the use of palbociclib in combination with aromatase inhibitors as first-line treatment of HR+/HER2- metastatic breast cancer. Clinical Oncology. 2024; doi:10.32471/clinicaloncology.2663-466X.49-1.30186.

63. Rasmy L, Wu Y, Wang N, et al. Med-BERT: pretrained contextualized embeddings on structured electronic health records for disease prediction. 2021 Jul 21. doi:10.48550/arXiv.2107.10239.
64. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med. 2020;26:44–56. doi:10.1038/s41591-020-1089-9.
65. National Academies of Sciences, Engineering, and Medicine. Proposed revisions to the common rule for the protection of human subjects in the behavioral and social sciences. Washington, DC: National Academies Press; 2014.
66. Ternopir C, Rice NM, Allard S, et al. Data sharing, management, use, and reuse: practices and perceptions of scientists worldwide. PLoS One. 2019.
67. OpenAI. ChatGPT (GPT-5.2). Large language model used for structuring, synthesis, and citation mapping of existing scientific literature. OpenAI, 2026.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the authors at:

Yuliia Cherednichenko
Email: Yuliia.Cherednichenko@intego-group.com

Tetiana Skorokhod
Email: Tetiana.Skorokhod@intego-group.com

Intego Group
35B Borychiv Tik Street
Kyiv 04070, Ukraine
Web: www.intego-group.com