Real-World Data Reliability and Integrity: Ensuring Real-World Data is Fit for Use for Regulatory Submissions

phuse.global

Contents

Revision History

Version	Date	Summary
V1	11 September 2026	

1. Introduction

The PHUSE Real World Evidence project 'Quality and Reusability of Real-World Data' (1) focuses on the usability of real-world data sources for regulatory submission purposes. To assess the fit for use of this data, we need to assess the relevance and the reliability of the data.

For the purpose of their guidance documents, the FDA defines real-world data (RWD) as data relating to patient health status and/or the delivery of healthcare routinely collected from a variety of sources (2).

This white paper aligns with the FDA definition and distinguishes the following types of data sources:

- Electronic health records (EHRs) contain patient encounters that typically encompass essential administrative and clinical data, such as demographics, progress notes, medical issues, prescribed medications, vital signs, medical history, vaccination records, laboratory results, and radiology reports. These comprehensive records offer a longitudinal perspective on patients' medical histories.

- Health insurance claims are based on the EHR data and include the financial aspects of healthcare. This may encompass information on indications, performed assessments, and procedures.

- Patient generated health data (PGHD) may include data from digital health technologies (DHTs) (e.g. commercial wearables and apps), social media and open surveys. Social media and patient surveys capture subjective insights, offering a unique perspective often absent in traditional research settings. DHTs bring real-time physiological data into the equation, providing a continuous stream of information.

- Patient registries serve as organised systems that collect uniform data on a population defined by a particular disease, condition or exposure, and are followed over time. The data in registries might be partly based on EHRs and additional collected information.

In essence, real-world data offers a real-life perspective on healthcare outcomes. If used appropriately, it can help us understand not only treatment effectiveness but also the wider societal and environmental factors affecting health and well-being.

This white paper provides an overview of specific requirements for fitness for use of real-world data that align with applicable guidance and regulations. It addresses the relevant regulatory guidance, the data provider assessments needed to comply with these guides, and the technical requirements for tools used to ensure the relevance and reliability of the real-world data for regulatory purposes.

2. Navigating Regulatory Guidance and Frameworks To Assess Fitness for Use of Real-World Data

Several regulatory guidance documents reference the quality of real-world data (RWD) and its suitability for regulatory submissions. These documents provide non-binding recommendations and frameworks to address data quality, study design and regulatory compliance. The RWD guidance emphasises two critical factors for ensuring RWD is useful in supporting regulatory decision-making for drugs and biological products: data relevance and reliability.

When both aspects are adequately addressed, RWD can serve as a valuable complement to traditional clinical trial data, enabling regulators to make informed decisions that reflect the realities of everyday clinical practice and patient experiences. The guidance collectively underscores the importance of data that is of sufficient data quality, well documented and contextually appropriate to advance healthcare and regulatory decision-making. In short, the data is then fit for use in regulatory trials.

The project group reviewed and discussed the following key guidance relevant to the fitness for use assessment of real-world data:

1. **Considerations for the Use of RWD and RWE to Support Regulatory Decision-Making for Drug and Biological Products** (FDA, August 2023) (3): Emphasises data quality and study design, focusing on RWE applications across both pre-market and post-market settings.

2. **Data Quality Framework for EU Medicines Regulation** (EMA, October 2023) (4): Encourages cross-stakeholder collaboration to ensure data meets regulatory standards.

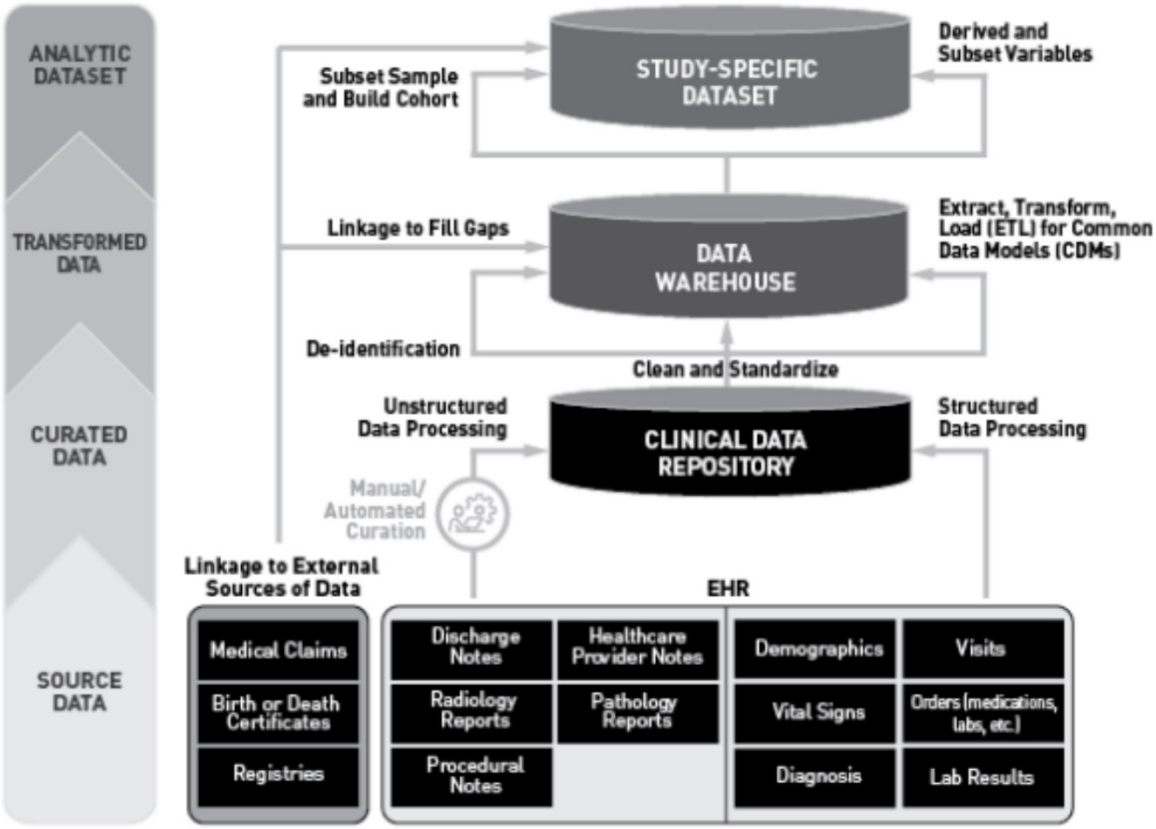
3. **Real-World Data: Assessing Registries to Support Regulatory Decision-Making for Drug and Biological Products** (FDA, December 2023) (5): Highlights how well-designed registries can enhance multiple studies, particularly for small or rare populations.

4. **Real-World Data: Assessing Electronic Health Records and Medical Claims Data to Support Regulatory Decision-Making for Drug and Biological Products** (FDA, July 2024) (2): Outlines strategies for data selection, validation, and handling missing data to meet reliability requirements.

The implications of some of this guidance are discussed in blog posts issued by our project group (6), (7) and in a poster presented at the PHUSE US Connect in 2024 (8).

2.1 The Lifecycle of RWD: From Source to Regulatory Submission

Figure 1: Illustrative Example of the Life Cycle of EHR Data[24]



To generate scientific real-world evidence using RWD, the FDA recommends establishing detailed, iterative and ongoing processes to account for the information needed to review and reproduce all the results.

Figure 1 from the FDA guidance that focuses on assessing EHR and medical claims data (2) to support regulatory decision-making illustrates the process of initial data capture to the derivation of a final analysis dataset by dividing it into five major phases:

- **Data accrual from the original source data**
- **Curation of data** to the clinical data repository
- **Transformation** and de-identification of data where necessary
- Creation of a **data warehouse**
- Production of a study-specific **dataset for analysis**

As stated in the guidance, the concept of the data life cycle illustrates the iterative nature of the process for examining the quality of the data. The reliability and integrity of the data should be checked and ensured at each step of the process.

In general, the reliability of the data, as it culminates into the final analysis dataset, is assured by characterising and, where possible, quantifying the **accuracy** and **completeness** of the data across the life cycle while maintaining **traceability** throughout.

To satisfy these requirements, analysis documentation submitted to the FDA before initiating data analysis should specify the procedures for preserving traceability. This includes describing the systems and processes for tracking the origin and transformations of data elements. At the time of submission, the supporting documentation (e.g. metadata, audit trails, coding system version histories, and access controls) enables the reconstruction of the data at each phase.

The FDA also recommends developing a Quality Assurance/ Quality Control (QA/QC) plan to describe how sponsors will address issues and outline the tools and methods for assessing the data – data quality checks, cleaning, and monitoring at every phase in the cycle.

2.2 Consistent Themes for Ensuring RWD Reliability, Relevance and Integrity

In their guidance, regulators frequently refer to the reliability, relevance and integrity of RWD to describe fit-for-use RWE in regulatory submissions. While the interpretation of these terms currently varies (9), we will identify and discuss the consistent themes.

2.2.1 Understanding the Data Collection Process(es)

Understanding the data collection process(es) is central to ensuring the reliability and relevance of RWD sources. Challenges such as bias and issues with representativeness

stem from the data collection process, especially in registries, which may fail to capture a fully representative sample of the target population. For instance, patients with more severe diseases are more likely to be overrepresented, leading to potential selection bias that can undermine internal validity. To enhance external validity and ensure the study's findings can be generalised, registry recruitment and retention strategies must mitigate these biases and more accurately reflect the broader population. These considerations align with the **FDA's Registry Guidance (December 2023)** (5), which emphasises the importance of understanding participant selection processes and the representativeness of registry and non-registry populations alike.

Integration of data from multiple sources, such as registries, EHRs and claims systems, presents additional challenges. Moreover, the linkage of data between data sources should be validated. This means ensuring accurate patient-level linkages and applying predefined rules for consistency, considering any missing data as well as multiple records of the same person at different sites, which can result in overcounts of a particular data measure or, if some site records are not available, in a collection of patient histories that reflect only a fraction of the patient's total healthcare history. It is also worth noting that these linkages may or may not be applicable to all participants within a data source. Some linkages may be jurisdiction-dependent or dependent on other characteristics of the participants, such as stage of disease. Clear documentation of the entire linkage process will be critical in understanding what inferences can be made from analyses of this data. The **FDA's EHR and Claims Data Guidance (July 2024)** (2) recommends documenting the type of curation performed to address duplication or fragmentation issues and documenting approaches taken to address issues that cannot be fully rectified by curation.

2.2.2 Informed Consent and Use of Personally Identifiable Information (PII)

Informed consent is fundamental for registries and other RWD sources, as emphasised in key regulatory guidance. Initial consent may allow future data use, but reconsent can introduce challenges when required later in a study. If only some participants agree to reconsent, this may introduce bias and affect the representativeness of the study population.

If data is pivotal to regulatory decisions, then quality assurance and control processes should be in place. Regional differences such as privacy regulations may complicate this process. De-identification, which removes or masks personally identifiable information (PII), is often used to address privacy concerns and enable secondary use of data. We will discuss this in section 5.

2.2.3 Feasibility Evaluation of RWD

Feasibility evaluations determine whether the selected RWD source aligns with the study objectives and captures key variables such as outcomes, exposures and covariates. The evaluation of relevant data sources or databases is an important step in the design of a study (3).

For example, when using real-world data to evaluate the effectiveness of a treatment, sponsors should confirm that the selected data source consistently includes key variables such as patient demographics, comorbidities, treatment histories and relevant outcomes such as hospitalisation or mortality rates. If the data lacks completeness (e.g. missing treatment details) or is inconsistent (e.g. different sites using varying definitions for a variable), it may lead to inaccurate conclusions. The assessment of the feasibility of data sources is addressed in section 4.

The FDA suggests upfront use of quantitative approaches (e.g. quantitative bias analyses) to get a good understanding of how misclassification of important variables (i.e. exposure/outcome/covariate values) may impact on study findings (2). Publications of these kinds of bias analyses are available (10) (11). In addition, the draft EMA Data Quality Framework for real-world data includes an extensive list of metrics that check the accuracy, precision, traceability, completeness, coverage, coherence and timeliness of a real-world data source (12). In summary, these evaluations are important in ensuring sponsors select data sources that are fit for use.

2.2.4 Pre-Specification of Data Selections and Analyses

Data sources, variable definitions and analyses should be predefined as indicated in the FDA EHR guidance (2). The guidance recommends pre-specifications of the inclusion and exclusion criteria for patient selection, the definitions for primary and secondary endpoints, and the statistical methods for analyses. The guidance emphasises the importance of outlining all derived variables, linking processes, and documenting potential biases before data is analysed. The selection of data sources or databases, as well as the conduct of specific analyses, should be free from bias towards a predetermined conclusion, providing regulators with the confidence to support regulatory decision-making (3). Pre-specification minimises post hoc adjustments and ensures reproducibility and consistency in regulatory submissions.

2.2.5 Early Engagement Across Stakeholders

When a regulatory decision is sought, the sponsors should engage with the regulators before conducting the study to agree on all essential elements of study design, analysis, conduct and reporting as indicated in the FDA RWD and RWE guidance (3). Early engagement reduces the risk of misalignment and addresses potential challenges. In this guidance, the FDA provides examples of how sponsors should communicate with regulators during study design, such as discussing the feasibility of linking multiple datasets, ensuring traceability of patient-level data, and pre-defining how missing data will be addressed. When analyses of RWD are used as evidence to support a drug's effectiveness or safety, sponsors should be prepared to make the patient-level data and source records (including the metadata) readily available for regulatory review. Within the sponsor team, this requirement should be made clear upfront. When working with third-party data providers, the sponsor should, in the development phase of the study, explore whether the above is feasible.

The challenges in this respect were presented at our Working Group Community Forum in January 2024 (13).

2.2.6 Governance and Data Quality Management

Governance forms the backbone of regulatory compliance, data integrity and technical reliability in RWD management. Not only is strong governance a regulatory necessity, but it also enhances operational efficiency, reduces compliance risks, and boosts the credibility of real-world data in clinical and regulatory decision-making.

Sponsors bear the responsibility of adhering to regulatory standards and guidance governing the use of RWE in drug development.

Gaps in governance can increase the risk of fragmented records, inaccessible data, and analyses that may produce biased conclusions and unreliable insights. A lack of structured oversight can make it challenging for organisations to verify data provenance, demonstrate regulatory compliance, and maintain accuracy across multiple transformations. Both the **FDA RWD and RWE Guidance** (3) and the **EMA Data Quality Framework** (4) stress the importance of formal governance practices that safeguard data integrity from collection through to regulatory submission.

As presented from an FDA perspective during our Community Forum in January 2024 (13), failure in governance can have very real consequences. In one regulatory review presented at this webinar, it was indicated that the lack of audit trail in the data collection process complicated the review. This oversight delayed their submission by months because additional validation was required.

Governance has many different aspects, as detailed below:

Documentation and Document Management

Complete and traceable documentation is important for data reliability and the integrity of study results. For instance, the **FDA RWD and RWE Guidance** (3) expects tracking data sources, transformation methods and verification steps. Similarly, the **EMA Data Quality Framework** (4) reinforces structured traceability to ensure datasets used in regulatory submissions remain linked to their original sources. Retention policies also align with regulatory expectations by preserving records for the required duration to safeguard essential validation data.

Contracts and Agreements

Contracts are critical in defining responsibilities between sponsors, data providers, and research institutions. Poorly structured agreements have, in some cases, led to sponsors losing access to crucial datasets because retention and verification rights were unclear. For example, the **FDA Registry Guidance (December 2023)** (5) indicates that sponsors seeking to use registry data to support a drug's effectiveness and/or safety in a marketing application should ensure that patient-level data are provided to FDA in accordance with applicable legal and regulatory requirements. Patient-level data in this case are essentially the SDTM and ADaM datasets that the sponsor submits to the NDA/BLA. The **FDA EHR and Claims Guidance** (2) on the other hand is more focussed on the data source verifiability. It indicates that the traceability of data coming from EHRs and medical claims data should be addressed to ensure the completeness and accuracy of study data. This includes the traceability of data elements to allow tracking of these elements back to their respective points of origin.

Robust contractual oversight is essential. Weak management in this area can lead to incomplete or unverifiable datasets, increasing the risk of regulatory delays. Good governance involves enforcing third-party compliance by requiring data providers to meet obligations regarding data traceability, audit trails and verification procedures.

Standard Operating Procedures (SOPs)

SOPs lay out the procedural framework for governance by standardising workflows for data extraction, transformation and validation. When these procedures are applied inconsistently, discrepancies can emerge across teams and systems, jeopardising data integrity.

Active Compliance Monitoring

Governance is not a one-time setup – it's an ongoing process. Active, real-time monitoring is crucial to validate adherence to policies and regulatory expectations. Continuous tracking helps prevent data discrepancies from propagating throughout the RWD life cycle. The **FDA RWD and RWE Guidance (August 2023)** (3) calls for systems that:

- Log modifications
- Document discrepancies.

Likewise, the **EMA Data Quality Framework (October 2023)** (4) requires validation methods such as temporal plausibility checks and cross-referencing with external comparators to confirm data accuracy. Both internal and external audits play a key role in detecting governance weaknesses early, preventing issues from escalating into significant regulatory findings. In the absence of active monitoring, governance issues may only become apparent during submission, delaying approvals and reducing data credibility.

Data Management

Governance frameworks should address data management throughout the data life cycle, as described in the **FDA EHR and Claims Data Guidance (July 2024)** (2). The guidance emphasises documenting key processes such as defining variables, managing data transformations, and maintaining clear audit trails. For example, sponsors are advised to create data dictionaries that include variable definitions, permissible values, and transformation rules. These dictionaries help track how data is processed and modified, supporting consistency and traceability from initial accrual to the final analytic dataset.

The guidance also highlights the importance of metadata and audit trails for regulatory review. Metadata documents the processes applied to data at each stage, such as ETL (extract, transform, load) operations and linkage methods. Missing data presents a significant challenge in using registry data, EHR and claims data for regulatory purposes, as outlined in the corresponding FDA guidance documents (2) (5). Both documents, as well as the EMA Data Quality Framework (4), indicate that missingness can have an effect on representativeness of the data for a population. The reason for missingness and drop-

out of patients should be well understood and handled when selecting and using the data for regulatory purposes. See section 4.1.

3. Assessing Data Source Feasibility

The FDA acknowledges that evaluating relevant data sources or databases is a critical step in study design and in assessing a study's feasibility (3). As outlined in the TransCelerate white paper on engaging with health authorities on real-world evidence (14), first, it is crucial to provide a strong justification for selecting the proposed data source(s) for the study to ensure final alignment with health authorities (HAs) on the protocol and statistical analysis plan (SAP). Second, such evaluations of data sources or databases for feasibility allow both the sponsor and the FDA to determine whether the data source is fit for use to address the research question and to estimate the statistical precision of the potential study, without the need to assess treatment arm outcomes.

3.1 Overall Feasibility Challenges

The PHUSE Quality and Reusability of Real World Data project team identified several challenges with assessing feasibility of a real-world data source. Although a data source may seem to be fit for use based on general characteristics, the data may prove to be less suited for the use than expected when it is reviewed to its full extent. This may cause real-world data studies to be terminated or not to start at all. The project group identified the following challenges:

Continuity of coverage

The FDA guidance for assessing Electronic Health Records and Medical Claims Data to Support Regulatory Decision-Making for Drug and Biological Products (2) indicates that the capture of a patient's healthcare information in a medical claims or EHR data source depends on continuity of coverage. A patient may change areas, caregivers or health insurance plans. During the PHUSE RWE Community Forum of October 2024, this issue was indeed noted from a data provider perspective (15). Depending on the source of the data, these continuity disruptions might influence the reliability of the observational data. Therefore, when selecting EHR data sources, continuity of care should be discussed, specifically whether patients receive all types of relevant healthcare services within the same health system or network of facilities that contribute to the same EHR data source. However, it can be difficult to assess whether a patient is continuously covered since it is not recorded as such and therefore may need to be inferred based on a proxy. This proxy could be, for example, based on expected visits and drug supplies.

It is important to understand and evaluate the methods used by the source to assess continuity of coverage to be able to address any issues that may arise from undetected coverage disruptions. This is also addressed in the FDA guidance on assessing EHR and medical claims data, which indicates that these methods should be noted in the study protocol, including potential impacts on study validity (2).

Data reliability and integrity

As stated in the TransCelerate white paper, data quality (e.g. reliability and integrity) can often be a concern to health authorities (HAs) due to missing, incomplete or redundant entries occurring at the data collection or management stage (14). They indicate that the measures, procedures and documentation, including quality assessments and/or quality checks that the data service provider has in place, can be informative for these discussions with HAs, particularly for the protocol and the SAP. The FDA guidance addresses data traceability and quality as well throughout the processes of data accrual, curation, and incorporation of corresponding details into the final study-specific dataset (2). Section 5 of this white paper will dive deeper into the corresponding technical data lineage methods. Our project group also identified that checks for uniqueness of data elements need to be included for data reliability.

Accuracy and representativeness

The selected data source needs to include an accurate and representative sample of patients who are relevant for the study's scientific question of interest. The EMA Data Quality Framework (4) emphasises the need to demonstrate how well the data reflects real-world conditions. In this framework, representativeness is defined as ensuring the data accurately captures both the patients and the characteristics of the population it is intended to represent. When used for regulatory purposes, we need to show that representativeness issues are not causing any bias in the study analysis results. For example, when estimating a treatment effect, representativeness is important to mitigate any potential treatment effect heterogeneity. Respective effect modifiers should be included in the data and relevant values of these variables represented. Poor representativeness on other factors that don't influence the treatment effect are less relevant and can be mitigated. The FDA guidance indicates it is important to identify that the data sources cover all populations relevant to the study question (2).

Data standards

Patient care facilities and sites use information systems that are usually dedicated to aiding their daily clinical caregiver practice. They will use their own data structures and have their own data entry capabilities. This affects what is entered and how it is stored. Common data models (CDMs) standardise this data into a common format for interoperability across all sources providing data. Standardisation can be managed and performed at different steps in the data flow process, which is also indicated in the FDA guidance (2). One option frequently used by EHR data providers is standardising to FHIR exchange format (16). However, it is important to recognise that when healthcare data is transformed to such a standard, the resulting dataset represents information that has already been selected and curated from the original records. It may include interpretations and mismatches and may not fully capture all the information recorded by the healthcare providers.

For optimal insights in the data and potential biases, it is important to understand what source information systems were used and what mapping and transformation actions have taken place before the data is assessed for feasibility.

3.2 Specific Challenges

Required Data Elements

As indicated in the FDA guidance (2), as well as by Gatto et al. (17) in their proposed process, it's important to establish a study design before conducting data feasibility assessments. This will identify the follow-up time and data elements needed to address the study objectives, select the appropriate/relevant patient population and include any potential confounders and safety variables to adjust for or summarise in the study. A data source that is fit for use for one study might not be for another. Availability and percentage of missingness in any of these data elements need to be assessed, as well as the follow-up time for patients.

Regional Differences

The FDA guidance (2) confirms there are differences in the practice of medicine around the world and between healthcare systems that may affect the relevance of the data source to the study question. This could be differences in diagnosis practice, medication tiering, patient handling, and even in disease characteristics. Therefore, it is important to know the regional healthcare and collection practice and, if possible, conduct bias assessments based on the available information. If applicable, imputations for missing data need to reflect the corresponding regional practice. Privacy regulations may vary and should therefore be addressed for each region.

Sample Data

In some cases, data providers may issue sample data to show what information is available from their source. However, members of our project team identified that this was not always representative of the actual data they received. The sample data might only include best cases and not indicate data issues such as missingness and representativeness.

Time Stamping

Audit trails are not always available for EHR source data. That means that over time entries made may change while unnoticed. For example, if an initial indication is found to be incorrect, the source data may be updated to reflect the correct indication. This can lead to discrepancies in the data that need to be addressed. Therefore, it is important to consider the method of audit trailing or time stamping for the source data when assessing its feasibility.

3.3 Feasibility Assessment

A thorough feasibility assessment ensures the data can be used for regulatory purposes. Requirement and assessment tools are available for the fitness for use assessment, such as the 'Structured Process to Identify Fit for Purpose Data' (SPIFD2) framework (17), a framework defined by the UK National Institute for Health and Care Excellence (NICE) (18), an EMA list of aspects to cover for registries (19), and regulatory submission evaluations such as the TransCelerate white paper mentioned (14). We also visited this topic from both sponsor and regulator perspectives in our most recent Community Forums (19).

The SPIFD2 identifies the minimal criteria for valid operationalisation in RWD. It combines multiple aspects and identifies issues related to the required data elements. Although it provides a good basis, it doesn't address all the issues in detail. Therefore, our project group developed a 'Data Provider Request For Information' document (DPRFI), which focuses on the regulatory requirements and issues (see Appendix 1). This can be used as a template to assess the data feasibility of data sources for a specific study. The output from this form can be used to fill the SPIFD2 table and be addressed in the clinical study protocol and/or alignment with regulatory health authorities.

The answers to the DPRFI will give a more complete picture of the data available and the usability of the data for the specific study than the tools currently available. However, we acknowledge that for a data provider it might be too much effort to generate the information needed before any commitment has been made. A stepwise contracting approach might help in this case, where an initial set of answers can lead to a deep dive based on an exploration agreement. Ultimately, it should be in the interest of all parties to show that the data is indeed useful for the intended purpose. This may also help in any follow-up use cases with the data. As stated in the TransCelerate white paper (14), based on diverse regulatory guidance, it is valuable to understand and include any regulatory/research use of the data source.

4. Technical Reliability of Real-World Data

Data from real-world data sources can be obtained via different routes from a data provider, a trusted third party and/or directly from the source. For governance, provenance and traceability, information is needed to substantiate how the data was collected, how correctness was ensured, how it was stored, how it was selected and how it was transformed during the whole data flow process from EHR to regulatory submission. This information should be secured as well as the metadata.

This section details suggestions for computerised systems to comply with these overall principles. They can be used to develop and/or evaluate a system that enables (part of) the digital data flow from EHR source to regulatory trial submission.

4.1 Data Anonymisation

Data anonymisation and pseudonymisation are two of the main approaches to removing personally identifiable information (PII) in real-world data and ensuring patient privacy without compromising data utility. While data pseudonymisation allows us to trace the information to the source, full data anonymisation removes any link to the source patient. For regulatory submission purposes, the full anonymisation approach is not acceptable since the corresponding data lineage cannot be checked and controlled (2). Instead, fully secured methods to both respect patient privacy and allow for traceability are increasingly used.

Maintaining detailed records of privacy measures, data usage policies, and monitoring processes is crucial for demonstrating compliance with regulations such as HIPAA (20) and GDPR

(21). In practice, this includes the database-specific privacy policies, re-identification risk assessments and computational transformations performed to make data available for evaluation. Although targeted to the SDTM submission standard, PHUSE published a white paper on the de-identification standard for SDTM 3.2 (25), which includes a Data Anonymisation and Risk Assessment Automation (DeID Toolkit) and PHUSE Good Transparency Practice (GTP) for assessing the de-identification risk.

Access limitations to source data based on privacy legislation is a critical hurdle in the use of real-world data for regulatory submission purposes. The informed consents associated with individual patient data are necessary for that access. These consents are either obtained at time of enrolment in the registry or when the data is coming from an EHR data source by the data provider. However, to comply with the privacy regulations, the data needs to stay de-identified for all the other processes related to the patient data.

There are several options in pseudonymisation to link the de-identified data back to the source:

- Via encryption: If patient information is encrypted via a fixed protocol, starting at the source, the matching encryption key can be linked to the de-identified data.
- Tokenisation: Patient information is linked to a token and transferred anonymously. Christine Rossin presented an example during our RWE Community Forum of October 2023 (22).

Those methods can be applied by a data source, a trusted third party (TTP), or the data provider.

In the data lineage from real-world data source to submission, the steps taken to de-identify the data as well as to exclude patients from the final database based on re-identification risk, needs to be precisely detailed to allow for accountability, transparency and traceability. This transparency is vital for audits and regulatory reviews, assuring that patient data is managed responsibly.

Excluding patients from a database based on privacy policies may cause unidentified bias. Therefore, a sensitivity analysis is needed to confirm that excluding these patients does not influence the study's analysis results.

4.2 Data Lineage

Both the U.S. Food and Drug Administration (FDA) and the European Medicines Agency (EMA) recognise the importance of data lineage for data quality and regulatory compliance throughout the data life cycle. The EMA defines data lineage as the presentation of how data originated and underwent processing before reaching its current form (4). Accurate and reproducible documentation of this journey of data, including data transformations and modifications, reduces the risk of corruption and misinterpretation and improves the value of real-world evidence (RWE) (3) (4).

Verifiable data integrity and unified governance controls across organisations and systems are critical to ensuring traceability and transparency for RWD reliability. Novel techniques play a pivotal role in creating a secure, tamper-evident record of system, code and codified human data interactions. They make sure each system serves as a deterministic source of truth, providing real-time documentation of all data transactions for data quality.

It is important to identify positive lineage, i.e. elements used in the analysis dataset, and negative lineage, i.e. elements not used in subsequent processing (CDISC, 2024). Using standardised data formats, such as CDISC, facilitates a more auditable data lineage (2). To evaluate lineage, it is essential to trace elements from the analysis dataset to the source. This ensures the data has been accurately captured and recorded. Conversely, forward checking verifies data completeness from source to output.

Maintaining the lineage of patient-level data elements in submissions can be challenging, especially when patient data is spread across multiple, unrelated databases. To address these complexities, regulatory agencies provide specific considerations for data lineage. The FDA's Real-World Data (RWD) and Real-World Evidence (RWE) Guidance (3) emphasises the necessity for sponsors to maintain audit trails and establish agreements with third parties for traceability. It is crucial for sponsors to avoid relying on external links and to allow direct access to patient-level data within controlled environments. This includes using SDTM variables such as External File (XFN) and External File Vendor Name (NAM), where the latter is accompanied by a two-character domain code. In addition, the European Medicines Agency (EMA) Data Quality Framework underscores the significance of temporal plausibility and adherence to expected data properties (4). The FDA Registry Guidance also mandates traceability for data elements sourced from registries and linked datasets (5). The FDA's EHR and Claims Guidance highlights the importance of maintaining patient-level traceability throughout the processes of data accrual, curation and transformation for compliance with regulatory standards (2).

To summarise, sponsors are encouraged to establish lineage practices to track data flow and validate transformations, as indicated in the FDA and EMA guidance. This depends on thorough documentation, systematic validation, and adherence to quality standards throughout the entire data life cycle. To establish verifiable data lineage back to source, organisations should contemporaneously document each action across the data life cycle. This documentation process should be automated, and connected to downstream use, and compliance should be monitored in real time. By digitally signing each action with unique system or user keys, organisations provide seamless backward traceability, enabling auditors to verify the data source and process lineage. Automation and visualisations are very helpful in showing the data lineage in a comprehensible form. New methodologies, for example cryptography, can ensure the immutability of this information.

While encryption protects data at rest and in transit, data can be vulnerable during processing when it's being actively used by applications. Confidential computing addresses this vulnerability by protecting data and code during execution at the hardware level. This involves using processor-based technologies that create secure, isolated environments known as trusted execution environments (TEEs). Confidential computing

provides the highest level of privacy and security for sensitive workflows, creating secure enclaves where data and code are protected during execution. This hardware-based security can trigger automated alerts for any unauthorised access attempts, supporting incident logging and response.

By implementing confidential computing solutions, organisations can protect sensitive patient data and proprietary code even during computation. These technologies prevent unauthorised access by isolating data and code within hardware-enforced boundaries, effectively safeguarding against threats that might exploit vulnerabilities during runtime. This level of security is crucial, as sensitive data and code could otherwise be exposed in system memory or processor logs.

4.3 Reliable Data Abstraction from Unstructured Data

Extracting insights from unstructured real-world data is critical for improving the quality and usability of healthcare datasets. Natural language processing (NLP) enables automated data abstraction, which is particularly effective for enriching structured datasets with valuable clinical information derived from clinician notes, pathology reports, and other free-text sources. Core applications of NLP include data enrichment, in which relevant clinical details are extracted to augment structured records, and categorisation and standardisation, wherein extracted information is mapped to established medical terminologies such as SNOMED CT, LOINC, ICD10 and RxNorm for facilitating semantic interoperability across systems.

However, integrating AI-based NLP methods into clinical data pipelines introduces regulatory and methodological complexity. While AI offers the potential for more rapid processing of unstructured healthcare data, it also introduces variability and uncertainty. The FDA notes these methods require significant human-aided curation and decision-making, injecting an additional level of data variability and quality considerations into the final study-specific dataset. Protocols should clearly describe the source data used, the type of algorithm (e.g. supervised or unsupervised), the assumptions and parameters involved, and the procedures for consistency in data extraction. Validation of metrics such as sensitivity, specificity and predictive values should be provided to demonstrate the accuracy and reliability of the NLP-extracted data used in study endpoints and key variables. These considerations ensure transparency, reproducibility and fitness for use of AI-derived real-world data in support of clinical and regulatory decision-making (2).

4.3.1 Validation and Traceability of the NLP Process for Unstructured Data

Ensuring the accuracy and reliability of NLP-driven data extraction requires a rigorous validation framework. Validation typically begins with comparison against a manually curated ground truth dataset, where model performance is evaluated using metrics such as precision, recall and F1 score (23). Interrater agreement measured by Cohen's kappa or Fleiss's kappa assesses consistency among human annotators during annotation phases. Clinical validation by subject matter experts (SMEs) confirms that the terms and clinical entities extracted by the NLP system are contextually correct and relevant to care settings (2). Error analysis and bias assessment are important for identifying and mitigating misclassifications and potential disparities in model performance across diverse populations. To support ongoing reliability, longitudinal consistency checks can verify that NLP models produce stable and reproducible results across evolving datasets.

Maintaining traceability throughout the NLP pipeline is essential for regulatory compliance and auditability. According to FDA guidance, sponsors are expected to document the assumptions used, and the algorithms applied, to define or extract relevant data elements, and to ensure reproducibility through clear versioning and methodological transparency (2). Data provenance tracking is a fundamental part of this traceability, allowing teams to record the entire data flow from unstructured input to structured output. Similarly, version control mechanisms for NLP models enable tracking changes in training data, model architecture, algorithm updates, and parameter configurations (2).

Retaining intermediate outputs such as tokenised text, extracted entities and transformation steps is also vital for stepwise validation and debugging. Audit logs and metadata capture (e.g. time stamps, system configurations, execution parameters) are essential for reconstructing processing history and supporting inspection or regulatory review (2). Finally, explainability techniques such as SHAP (SHapley Additive exPlanations) and attention visualisations in transformer-based models such as BERT provide critical insight into how the NLP model arrives at specific decisions, which improves transparency and user trust in clinical environments.

The following tools and frameworks support traceability across the NLP pipeline, from data ingestion to model deployment.

1. Data Versioning

Effective traceability starts with robust data versioning. Each dataset used, whether raw, clean or transformed, should be versioned to track changes over time. This ensures reproducibility of training and evaluation, especially when data sources are frequently updated. Tools such as DVC (Data Version Control) and Pachyderm can automate versioning and associate each data version with specific model runs or processing stages, allowing teams to revert, audit or compare outcomes across data iterations. The choice of tools depends on the storage type. DVC suits local or hybrid setups with Git, while LakeFS and Quilt are better for cloud object storage. Aligning versioning strategy with storage infrastructure is key to scalable, end-to-end traceability.

2. Data Lineage Tracking

Data lineage involves documenting the full life cycle of data from its original source to its final structured form, capturing steps such as cleaning, tokenisation and feature transformation. This traceability allows model outputs to be linked back to their inputs, supporting debugging, reproducibility and audit readiness. Kedro, Apache Airflow and ZenML enable step-by-step tracking in NLP pipelines, while DataHub and Marquez offer metadata and lineage visualisation across data and ML workflows, which is especially valuable when working with

clinical or sensitive text data.

3. Model Versioning

Each trained model should be treated as a distinct artifact with its own version, linked to the specific data, parameters and training environment used. By maintaining a model registry that tracks these elements, teams can reproduce past results, roll back to earlier models, and monitor performance trends. Platforms such as MLflow and Weights & Biases offer structured tools for registering and managing model versions, ensuring full life cycle traceability.

4. Pipeline Modularity

Designing NLP pipelines with modular components enhances transparency and traceability. Each stage – preprocessing, feature extraction, training, validation and inference – should be separated and clearly documented. Modular pipelines built with tools such as Kedro or Apache Airflow support reusable components, consistent data flow and easier debugging. They also allow for independent validation of each step, which is essential for complex workflows.

5. Experiment Logging

All training and evaluation experiments should capture model hyperparameters, performance metrics, training duration, hardware used, and random seeds. Logging enables reproducibility and makes it possible to understand why one model version outperforms another. MLflow and Weights & Biases provide structured dashboards and APIs for capturing this metadata across multiple runs, which also aids in collaboration and decision-making.

6. Audit Logging

To enable regulatory compliance and internal governance, systems should maintain audit logs that record when, how and by whom a pipeline was executed. This includes tracking configuration settings, runtime parameters, system environments and execution time stamps. Audit logs provide a detailed processing history that can be reconstructed later, which is vital in regulated environments such as healthcare and finance. MLflow, ZenML and Apache Airflow support audit logging by capturing metadata, execution traces and pipeline step logs. For NLP-specific workflows, platforms such as Weights & Biases and Neptune.ai offer detailed logging of training runs, dataset changes and model artifacts, for end-to-end traceability and compliance.

7. Explainability Integration

Incorporating explainability into the NLP pipeline helps trace how models arrive at specific decisions. Techniques such as SHAP (SHapley Additive exPlanations) or LIME highlight which features contributed to a prediction, making the model's behaviour more transparent. In transformer-based models such as BERT, attention visualisation can reveal how the model processes contextual information. Logging these explanations as part of inference outputs helps build trust with stakeholders and ensures accountability.

8. Intermediate Output Storage

Storing intermediate outputs at various stages of the pipeline – such as tokenised text, embeddings, extracted entities, and cleaned data – allows for stepwise validation and debugging. This enables analysts to verify correctness at each stage without rerunning the entire pipeline and ensures transparency in how the final output was derived from the original input. These artifacts should be versioned and stored alongside results.

9. Data and Model Monitoring

Ongoing monitoring of data and model behaviour is critical to detect issues such as data drifting, concept drift or performance decay. For instance, if the data distribution changes due to new patient populations or documentation styles, the model's effectiveness may decline. Evidently AI and whylogs can automatically detect these shifts and alert teams to retrain or revalidate the model. Monitoring ensures models remain trustworthy over time.

10. Documentation and Governance

Comprehensive documentation of all aspects of the NLP pipeline including data assumptions, model logic, evaluation criteria, and validation methods is essential for traceability. This documentation should be kept current and accessible to stakeholders, auditors and regulatory bodies. It supports explainability, facilitates onboarding of new team members, and aligns the pipeline with organisational policies and industry standards.

By integrating robust validation measures with transparent traceability mechanisms and selecting tools that align with the system architecture, data setup, and security requirements, organisations can make sure AI-driven NLP extraction from unstructured healthcare data meets high standards of quality, reliability and regulatory compliance. The tools mentioned are just a few among many available, and implementation should be tailored to each organisation's infrastructure, compliance needs and risk posture.

5. Discussion and Conclusion

Data quality should be sufficient to support decision-making and avoid misinterpretation that could undermine the credibility of trial results. This involves assessing the reliability and relevance of the data to the study objectives, implementing strategies to address potential challenges such as missing data, and establishing governance frameworks to manage data quality throughout its life cycle. Incorporating these practices into study design ensures the data collected supports robust and defensible regulatory submissions.
Efforts to achieve seamless interoperability and accurate data

linkages also enhance the utility of RWD. By addressing issues such as bias, consistency in data capture, and traceability of data elements, sponsors ensure the data they are using can answer the efficacy and safety questions for their compound/intervention of interest and meet regulatory expectations while making meaningful contributions to healthcare innovation.

The use of RWD in regulatory submissions presents both opportunities and challenges. Sponsors should engage with regulators early to align study design, data sources and statistical methods and avoid costly redesigns to meet regulatory expectations later. Supporting regulatory decision-making helps bring safe and effective treatments to patients in a timely and efficient manner.

6. Glossary

EHR (electronic health record)
An EHR is a digital version of a patient's medical record, created and maintained within healthcare systems. EHRs include structured data (e.g. demographics, diagnoses) and unstructured data (e.g. physician notes). EHRs are used for clinical care, billing and regulatory compliance. Common challenges include incomplete records, variability across healthcare systems, and the need for data curation when used for research or regulatory submissions.

Fitness for use
Suitability for a specific regulatory question.

Reliability
Reliability refers to how consistently and accurately the data reflects what it's intended to measure. This includes assessing **accuracy**, which evaluates discrepancies between the available data and source records, and **plausibility**, which considers whether the information logically aligns with expectations. Reliability also requires **traceability**, supported by audit trails and documentation that record the origin and processing of data throughout its life cycle. For instance, if data collection processes are inconsistent, or if variables critical to study objectives, such as treatment of exposure or outcomes, are incomplete or inadequately documented, the RWD may be considered unreliable.

Relevance
Relevance measures whether a dataset contains the data elements necessary to address a specific research question or regulatory objective. This involves evaluating whether the registry or data source captures the **key variables** – such as treatment of exposure, outcomes, and confounding variables – needed for analysis. It also considers whether the dataset aligns with the scope of the study population and whether it is comprehensive enough to represent the reality being studied. For example, a dataset must include the appropriate number and type of variables to support the study objectives, such as tracking long-term outcomes in a chronic disease study. Missing key variables or incomplete data collection over time may indicate that the dataset is less relevant for the intended purpose.

Registry
A registry is an organised system for collecting clinical and other data in a standardised format for a specific population, often defined by a disease, condition or drug exposure. Registries may be used for research purposes, such as providing external control arms or identifying participants for interventional studies. Key considerations include consistency in data capture and compliance with privacy regulations during data collection and management.

Medical claims
Medical claims data is records of healthcare services billed to payers. This data typically includes diagnostic codes, procedures and dates of service. While medical claims data is valuable for large-scale analyses, it has limitations, such as variability in coding practices and incomplete representation of clinical care. Careful evaluation is needed when using medical claims data for regulatory studies.

7. References

1. **PHUSE.** Real World Evidence Working Group. [Online] https://advance.hub.phuse.global/wiki/spaces/WEL/pages/26805467/Real+World+Evidence

2. **FDA CDER CBER OCE.** Real-World Data: Assessing Electronic Health Records and Medical Claims Data To Support Regulatory Decision-Making for Drug and Biological Products: Guidance for Industry. [Online] July 2024.

3. **FDA**. Considerations for the Use of Real-World Data and Real-World Evidence To Support Regulatory Decision-Making for Drug and Biological Products: Guidance for Industry. [Online] August 2023. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/considerations-use-real-world-data-and-real-world-evidence-support-regulatory-decision-making-drug

4. **EMA.** Data Quality Framework for EU medicines regulation. [Online] October 2023. https://www.ema.europa.eu/system/files/documents/regulatory-procedural-guideline/data-quality-framework-eu-medicines-regulation_en_1.pdf

5. **FDA, CDER, CBER, OCE.** Real-World Data: Assessing Registries To Support Regulatory Decision-Making for Drug and Biological Products: Guidance for Industry. fda.gov. [Online] December 2023. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/real-world-data-assessing-registries-support-regulatory-decision-making-drug-and-biological-products. FDA-2021-D-1146

6. **PHUSE RWE Working Group – Quality and Reusability of Real World Data project.** Considerations for the Use of Real-World Data and Real-World Evidence to Support Regulatory Decision-Making for Drug and Biological Products. [Online] June 2024. https://phuse.global/Communications/PHUSE_Blog/considerations-for-the-use-of-real-world-data-and-real-world-evidence-to-support-regulatory-decision-making-for-drug-and-biological-products

7. **PHUSE RWE Working Group – Quality and Reusability of Real World Data project.** RWD: The FDA's Guidance on Assessing Registry Data to Support Regulatory Decision-Making. [Online] July 2024. https://phuse.global/Communications/PHUSE_Blog/rwd-the-fda-s-guidance-on-assessing-registry-data-to-support-regulatory-decision-making

8. **PHUSE RWE Working Group – Quality and Reusability of Real World Data project.** Real World Data – What is Fit For Purpose? [Online] March 2024. https://phuse.s3.eu-central-1.amazonaws.com/Archive/2024/Connect/US/Bethesda/POS_PP25.pdf

9. *A Brief Report on Proposed Areas of International Harmonization of Real-World Evidence Relevance, Reliability and Quality Standards Among Medical Product Regulators.* **Nafie M, Parker VJ, McClellan M, Hendricks-Sturrup RM.** 2025, Pharmacoepidemiol Drug Saf., 34(3).

10. *Opioid-related Mortality in United States Death Certificate Data: A Quantitative Bias Analysis With Expert Elicitation of Bias Parameters.* **Goldsmith ES, Krebs EE, Ramirez MR, MacLehose RF.** 2023, Epidemiology, 34(3), pp. 421–429.

11. *Use of quantitative bias analysis to evaluate single-arm trials with real-world data external controls.* **Gray C, Ralphs E, Fox MP, Lash TL, Liu G, Kou TD, et al.** 2024, Pharmacoepidemiol Drug Saf., 33(5):e5796.

12. **EMA.** Data Quality Framework for EU medicines regulation: application to Real-World Data: Draft. [Online] September 2024. https://www.ema.europa.eu/en/documents/other/draft-data-quality-framework-eu-medicines-regulation-application-real-world-data_en.pdf

13. **PHUSE Community Forum.** RWD Quality Challenges – Different Perspectives. [Online] January 2024. https://phuse.s3.eu-central-1.amazonaws.com/Archive/2024/Working%20Group%20Events/Worldwide/Virtual/REC_WG01.mp4

14. **TransCelerate.** Preparing To Engage with Health Authorities on Real World Evidence for Medicines. [Online] November 2022. https://www.transceleratebiopharmainc.com/wp-content/uploads/2022/11/Preparing-To-Engage-With-Health-Authorities-On-RW-Evidence-for-Medicines_A-TC-Perspective-2.pdf

15. **Matthew Reynolds, Azimuth Research Group.** RWD Sources – How the Genesis of Your Data Determines What Questions You Can Answer. [Online] October 2024. https://phuse.s3.eu-central-1.amazonaws.com/Archive/2024/Working%20Group%20Events/Worldwide/Virtual/REC_WG07.mp4

16. **HL7.** HL7 FHIR. [Online] https://hl7.org/fhir/

17. *A Structured Process to Identify Fit-for-Purpose Study Design and Data to Generate Valid and Transparent Real-World Evidence for Regulatory Uses.* ***Gatto NM, Vititoe SE, Rubinstein E, Reynolds RF, Campbell UB.*** 2023, Clinical Pharmacology & Therapeutics, 113(6), pp. 1235–1239.

18. **National Institute for Health and Care Excellence.** *NICE real-world evidence framework.* June 2022.

19. **EMA.** Guideline on registry-based studies. [Online] September 2021. https://www.ema.europa.eu/en/guideline-registry-based-studies-scientific-guideline

20. **U.S. Department of Health and Human Services.** Health Information Privacy. [Online]

21. **European Union.** EUR-Lex Document 32016R0679. [Online] https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=celex%3A32016R0679

22. **PHUSE Community Forum.** Experiences with Adapting RWD to CDISC Submission Standards. [Online] October 2023. https://phuse.s3.eu-central-1.amazonaws.com/Archive/2023/Working%20Group%20Events/Worldwide/Virtual/REC_WG06.mp4

23. *Implementing Accuracy, Completeness, and Traceability for Data Reliability.* **Riskin DJ, Monda KL, Gagne JJ, Reynolds R, Garan AR, Dreyer N, Muntner P, Bradbury BD.** 2020, JAMA Netw Open., 8(3): e250128.

24. *Emulating Randomized Trials by Observational Database Studies: The RCT-DUPLICATE Initiative in COPD and Asthma.* **Suissa S, Schneeweiss S, Feldman WB, Tesfaye H, Wang S.** 2024, Am J Epidemiol, 194(5):1152–1159.

25. **Data Transparency.** *PHUSE Good Transparency Practice* [Online] August 2025. https://phuse.s3.eu-central-1.amazonaws.com/Deliverables/Data+Transparency/WP-080.pdf

8. Acknowledgements

Thank you to our project members for actively contributing to this white paper.

Ashwin Rai, Thermo Fisher

Berber Snoeijer, ClinLine

Elizabeth Merrall, Johnson & Johnson

James Joseph, EDA Clinical

Lihai Song, Merck

Furthermore, we would like to acknowledge the following project members for their contributions to the project and providing their input.

Benjamin Ackermann, Johnson & Johnson

Cheryl Grandinetti, FDA

D'dra Octaviano, Alnylam

Devaki Thavarajah

Ela Talpes

Matt Baldwin, Amgen

Mike Kolzet, EQTY Life Sciences

Shailesh Chavan, IG Health

Disclaimer: Because of the varied composition of our project group, the list might not be exhaustive. Thank you for all the contributions also from all the other project members.

Appendix 1: Data Provider Request for Information Form (DPRFI)

Note: text in yellow are only instructions and should be removed in final request form: Instructions for creating the RFI before sending it to the data provider are indicated in yellow.

1. Study-Specific Requirements

1.1 Inclusion and Exclusion Criteria

1. *Indicate specific criteria for the cohort and/or study before sending the RFI to the data provider: The RFI should include a list of criteria that the requested cohort should fulfil*

2. Please provide an overview of the number of patients fulfilling the criteria mentioned above, preferably illustrated by an attrition diagram.

1.2 Relevant Data Elements

The data elements required for the study, as described in the study synopsis, are listed below. Please describe the availability of each element and potential options for defining a proxy if not available.

Add data elements to the column requested by the sponsor based on the indication, objectives and endpoints and any potential confounders you like to include in the analyses. Indicate whether they are preferred or required for the data source to be considered.

Data element	Requested by the sponsor	Required/ preferred	Availability from the provider
Study indication variables			
Exposure/comparator variables			
Primary endpoint/outcome variables			
Secondary endpoint/outcome variables			
Exploratory endpoint/outcome variables			
Confounder variables			

See example RWE study emulating COPD IMPACT RCT by Suissa et al. (24)

2. Data Source Outline

2.1 Data Source Origin

Please provide a description of the origin of the data sources used to address the requirements of the study. This includes the collection setting and the original use of the data.

2.2 Data Governance

Please describe the data governance and corresponding process relevant for the data source(s) included.

2.3 Data Linkage and Synthesis Methods

If applicable, describe how multiple data sources used to address the requirements of the study are linked. Include a short description of the corresponding linkage accuracy, and indicate whether your data can be tokenised and combined with other data providers.

2.4 Applicable Data Standards

Please describe the standards applicable to the data sources (OMOP/ICD10/ATC, etc.). If data is combined from multiple data sources or study sites, indicate what standards are used per source/site.

Data model	
Medication	
Indications/Co-morbidities	
Laboratory assessments	
Procedures	
Other	

3. Data Quality and Usability

3.1 Completeness and Coverage

Please describe the following aspects for the cohort above with regards to the data source(s) indicated to create these cohorts.

Coverage: What countries and districts are covered and for what percentage of the total population?	
Active sites: How many active sites do you have available to support data collection for this project? Please provide their locations and capacities.	
Setting: What percentage of your patients come from a community care setting? What percentage of your patients come from an ac-ademic care setting?	
Missingness: What methods are used to handle missingness of the cohort inclusion and exclusion criteria?	
Period: What periods are covered by your data sources (i.e. the range from year to year) and, if periods are specified, for what percentage of the subjects is it applicable?	
Continuity of coverage: What methods are used to assess and/or assure data continuity of coverage of data for specific patients?	
Follow-up: What is the median database coverage period per subject before and after the expected index date*?	
Age: What is the age distribution (five-year age bands) of the subjects in your (linked) database?	

* Index date is based on the first time all eligibility criteria is met in the database.

3.2 Data Provenance and Traceability

Please describe the following aspects for the data source(s) indicated to create the cohorts.

Source information systems: How many dif-ferent information systems are used for data collection? If possible, provide system names and the percentage of the total database.	
Source audit trails: What methods are used to track changes of data in the source?	
Source bias: Is it possible to anonymously identify different information systems, prescribers and practitioners in the database to assess potential source bias?	
Source data mapping: What information can you provide regarding the ingestion, transformation and standardisation of source data relevant for this request?	
Local methods of diagnosis and preferred treatments for the disease: Describe and/or add links to local diagnosis and treatment protocols and programmes related to the study indication(s) in scope, if applicable.	
Local prescribing and use practices: Describe and/or add links to local prescribing and use practices related to the study indication(s) in scope, if applicable.	

3.3 Data Quality and Accuracy

Please describe the following aspects for the data source(s) indicated to create the cohorts.

Representativeness: Describe how the repre-sentativeness of the dataset can be tested and assured.	
Missingness of event data: Event data is data that is only recorded when it is present and not when it is absent (like medical events). Describe how the completeness of event data is tested and assured in the light of continuity of coverage. If applicable, have your mortality/progression metrics been validated?	
Data curation for deduplication of patients: Specify how testing for duplication of records and/or patients is done and how corresponding issues are resolved.	
Data lineage: Specify how data lineage from source to delivered dataset is ensured, including documentation and technical solutions to verify the traceability.	
Other quality measures: Describe any other quality measures you can provide for the dataset.	

3. Data Services and Delivery

3.1 Data Granularity

In what level of detail can you provide the data?

Format	Yes/No	Comments/Considerations
Patient level – based on informed consent		
Patient level – pseudonymised		
Patient level – anonymised		
Aggregated (Provide details on aggregation level possible*)		
Analysis results based on pre-aligned algorithms		
Analysis results only		

* If data delivery is only possible on an aggregated level, please substantiate how treatment information can be provided (constructed treatment periods or else) and how patients are grouped without losing information on overall timing effects.

3.2 Data Vendor Experience

Please provide details of any experience with the data sources used for regulatory studies in general and for the specific indication and topic of this RFI.

Do you have experience building prospective observational databases for partners?

If yes:

- How many studies did you build between 2021 and 2024?

- Provide a range of consent rates seen in your pros (1–100%).

- Provide a range of dropout rates seen in your prospective studies (1–100%).

3.3 Data Dictionaries and Sample Data

Would you be able to provide a data dictionary and/or sample dataset with mock data? Please give details of what, when and in what format this can be delivered.

3.4 Costs

Please provide details on your cost model with regards to the data sources specified in this RFI.