TrialIQ Preprocessing Framework: Structured Transformation of Clinical Trial Protocol Data for Predictive Modelling

Author: Francis Osei
Email: Francis@bayezian.com

1. Introduction

The integrity of any data-driven research in clinical science rests not merely upon the availability of data, but upon the careful curation, transformation, and contextual understanding of that data before analysis begins. This is particularly true in the context of clinical trial records, which are rich in content but often inconsistent in structure, heterogeneous in terminology, and incomplete in critical areas. Preprocessing, therefore, is not a trivial preparatory step, but a foundational phase of scientific inquiry, one that directly shapes the quality, reproducibility, and interpretability of downstream results.

In the present work, we approach preprocessing not as a background task, but as a formalised research stage in its own right. Drawing upon structured clinical trial records from a public registry, the study undertakes the systematic conversion of semi-structured protocol metadata into an analysis-ready form. This includes the standardisation of numerical and categorical fields, the linguistic parsing of eligibility criteria and outcome descriptions, and the application of controlled vocabularies to unify disparate trial records under coherent classification schemes.

Crucially, this report places emphasis on preserving clinical meaning while enhancing computational accessibility. Variables such as masking strategy, intervention type, or regulatory oversight, often inconsistently reported, are reconciled through rule-based logic and encoded in a manner that retains interpretive value. In parallel, key linguistic fields are distilled into numerical summaries without discarding their original text, thereby maintaining a dual path for quantitative and qualitative exploration.

The rationale for this work lies in the recognition that robust predictive modelling cannot be built atop unstable inputs. If features are ill-defined, values inconsistent, or missingness unchecked, then no algorithm, however sophisticated, can offer credible inference. Accordingly, the present study treats preprocessing as a matter of both methodological rigour and epistemic responsibility.

The sections that follow provide a detailed account of each preprocessing phase, with clear justifications, methodological transparency, and reflection on known limitations. In doing so, this work aims to serve as both a record of implementation and a contribution to best practice in the preparation of clinical trial data for predictive analysis.

2. Objectives

The overarching aim of this preprocessing phase is to convert raw clinical trial protocol data into a structured, consistent, and semantically meaningful format suitable for downstream analytical tasks. Given the heterogeneity and partial incompleteness of the source records, this process requires more than simple reformatting; it demands a deliberate, evidence-informed strategy for cleaning, standardising, and augmenting the original data.

To this end, the specific objectives of the preprocessing stage are as follows:

1. To normalise raw protocol fields
   Convert loosely structured attributes such as age ranges, masking strategies, trial phases, and Boolean flags into harmonised formats, correcting inconsistencies and enforcing standard representations across the dataset.
2. To extract high-value textual features
   Apply natural language processing techniques to eligibility criteria and outcome descriptions to generate token counts, sentence structures, and content indicators that capture the narrative complexity of trial protocols.
3. To encode categorical fields using interpretable groupings
   Group and transform trial design elements such as intervention types, sponsor classes, and randomisation strategies into consistent categorical variables, supporting both modelling fidelity and human interpretability.
4. To assign condition-level disease groupings via controlled ontology
   Employ biomedical vocabularies (e.g., MeSH) to map condition names to higher-order therapeutic categories, enabling domain-level stratification and reducing semantic drift across disease labels.
5. To calculate and impute derived variables
   Derive additional features, such as trial duration, from structured dates, and impute missing values where appropriate, ensuring analytical continuity without introducing artificial bias.
6. To retain fidelity to the original data while enabling robust modelling
   Preserve the original clinical semantics of each trial entry while transforming it into a form that can support statistical modelling, visualisation, and evidence synthesis.
7. To handle large datasets with computational efficiency
   Implement parallel processing where applicable, enabling the timely preprocessing of thousands of trials without compromising data quality or process transparency.

Taken together, these objectives reflect a commitment to methodological rigour, clinical sensibility, and analytical readiness. They also recognise preprocessing as an intellectual and technical undertaking that merits careful documentation, reflection, and reproducibility.

3. Source Data Overview

The dataset under consideration comprises structured protocol records retrieved from a public clinical trial registry. These records are intended to reflect the design and operational intent of interventional studies across a wide range of therapeutic areas and geographical contexts. Each trial entry includes a combination of structured fields (e.g., sponsor type, masking method, eligibility ranges) and unstructured or semi-structured text (e.g., outcome measures, eligibility criteria, summaries).

In principle, this registry offers a rich empirical resource. Yet in practice, the records vary widely in their completeness, consistency, and semantic clarity. Some trials are reported with exemplary precision; others omit key fields or use idiosyncratic terminology that resists easy interpretation. Certain fields are optional at the point of registration, and even mandatory fields are not always populated in a manner that supports comparability across entries.

Several key characteristics of the raw data deserve specific mention:

3.1 Structural Inconsistencies

While the registry imposes a nominal schema, the real-world application of that schema is uneven. For instance, masking strategies are variously reported as “Double,” “Double-blind,” “Blinded,” or occasionally misspelled. Sponsor types range from clearly defined terms (e.g., “Industry”) to ambiguous or missing labels. Age eligibility thresholds may appear in years, months, or weeks, often without units, and require transformation into a common numerical standard.

3.2 Missing or Ambiguous Values

Many trials omit one or more fields relevant to design characterisation. This includes, but is not limited to: the number of study arms, primary purpose, eligibility criteria text, or outcome measures. In some cases, dates are missing or inconsistently formatted, rendering duration calculations problematic without further imputation. Additionally, Boolean indicators such as “is FDA-regulated” are frequently encoded as free text (“Yes,” “No,” “True,” “None”) rather than binary values.

3.3 Semantic Heterogeneity in Condition Labels

The condition terms associated with each trial are often drawn from uncontrolled vocabularies, resulting in wide variation in spelling, granularity, and disease framing. For example, entries may list “Type 2 Diabetes,” “Diabetes Mellitus Type II,” or simply “T2DM”, all of which refer to the same underlying condition but appear distinct to a naïve parser. This necessitates the use of controlled biomedical ontologies to reclassify and consolidate such terms.

3.4 Repetition and Duplicates

Due to how records are retrieved (often using multiple disease-based queries), some trials appear more than once in the raw dataset. While these repetitions are generally exact duplicates, they must be explicitly removed to avoid bias in subsequent frequency counts or feature distributions.

This source landscape underscores the necessity of a thoughtful and rigorous preprocessing strategy. Without such intervention, the data, though voluminous, risks being analytically brittle, semantically incoherent, and prone to distortion when used for quantitative inference.

4. Preprocessing Workflow Summary

The preprocessing phase undertaken in this study comprises a sequence of methodical transformations designed to render raw clinical trial records analytically reliable. Rather than treating each field in isolation, the approach integrates multiple layers of logic, normalisation, and domain-informed inference to preserve the original intent of the trial while refining its structure and semantics for research purposes.

The workflow begins with the normalisation of raw input fields. This includes converting textual inconsistencies to a common standard, such as enforcing lowercase for categorical variables, trimming excess whitespace, and mapping semantically equivalent entries into a unified form. Fields like masking,

sponsor class, and intervention model are particularly susceptible to inconsistent labelling and thus require targeted regularisation.

Textual content, especially eligibility criteria and outcome descriptions, is subjected to tokenisation and sentence-level segmentation. This allows for the derivation of linguistic features such as token counts and structural complexity indicators, which are then retained as numerical variables for modelling. Importantly, the original text is preserved alongside these derived features to allow for future qualitative or language-based analysis.

Subsequently, categorical fields are encoded using a principled grouping strategy. For instance, sponsor classes are mapped into higher-order categories such as industry, government, or non-industry entities. Similarly, intervention types are summarised both by count and by composite archetype, such as whether a trial uses drugs exclusively, a combination of drugs and biologicals, or a more complex mix of device and behavioural interventions.

In parallel, condition terms listed within each trial are mapped to top-level disease categories using a controlled biomedical vocabulary. This step ensures that trials can be compared or aggregated at the level of therapeutic area, rather than being limited to their idiosyncratic condition labels.

Derived fields such as trial duration, calculated from registered start and end dates, are incorporated wherever possible. For entries where duration is missing or improperly formatted, lookup-based imputation is employed to recover values using matching records from more complete trials.

Throughout, all fields are preserved unless demonstrably uninformative or redundant. No variables are excluded arbitrarily, and pruning is deferred to later stages of modelling or analysis depending on relevance and missingness.

This workflow reflects a commitment to methodological clarity, clinical sensitivity, and analytical preparedness. It ensures that the resulting dataset is not only internally consistent but also structurally and semantically robust enough to support a range of exploratory and inferential techniques.

5. Field Normalisation and Standardisation

The initial stage of the preprocessing pipeline involves standardising the structure and representation of key trial fields. This process addresses inconsistencies in formatting, variation in data types, and the lack of coherence across seemingly identical values. All transformations are applied in a way that preserves the original information wherever possible while improving comparability across records.

Age ranges, when provided as text strings, are parsed into numerical values expressed in years. The normalisation logic supports conversions from weeks and months where units are detected, and assumes years when units are missing. Values that cannot be interpreted are set to null. If no valid age bounds are provided, defaults of 0 for the minimum and 100 for the maximum are assigned. This ensures that each trial record retains age eligibility parameters, even when they were not cleanly specified in the raw input.

Text fields such as masking, randomisation, intervention model, and sponsor class are converted to lowercase and stripped of surrounding whitespace. This standardisation ensures that trials using

syntactically different but semantically equivalent terms are treated consistently. These values are not categorised at this stage but are passed forward in a uniform format for subsequent encoding.

Study phase entries, when present as strings, are converted to uppercase. Entries reported as missing or non-standard are left unchanged. Similarly, date fields such as trial start and completion dates are parsed into date objects when possible. If both start and end dates are successfully parsed, the duration of the trial is computed in days and stored as a new variable. Records lacking one or both dates are flagged accordingly using a binary indicator that identifies missing duration data.

Boolean fields are handled cautiously. Values in fields such as "is_fda_regulated_drug", "healthy_volunteers", or "has_expanded_access" are converted to actual Boolean values if they are specified as “true” or “false” strings. If the values are neither Boolean nor missing, a placeholder value is retained, allowing for neutral representation without assuming intent.

Uninformative fields such as “oversight_authorities” are removed from the dataset entirely, as they are not required by any downstream components of the pipeline. All other fields are left intact, allowing maximum preservation of the original dataset for inspection or later filtering.

By applying these transformations in a deterministic and reproducible manner, the preprocessing routine ensures that every record is brought into a consistent format, free from superficial inconsistencies that could otherwise obscure analytical findings.

6. Eligibility and Outcome Text Features

In addition to structured fields, clinical trial records often contain unstructured narrative content, particularly within eligibility criteria and outcome descriptions. While such fields are not always amenable to direct numerical interpretation, they do carry meaningful signals about trial complexity, design intent, and reporting completeness. For this reason, the preprocessing phase includes the extraction of a small number of interpretable textual features from these sections.

The eligibility criteria text, when present, is tokenised to produce a simple count of the number of words or fragments it contains. This token count serves as a crude but informative proxy for the verbosity or complexity of the inclusion and exclusion logic underpinning participant selection. Trials with more elaborate or verbose criteria tend to score higher on this measure, although no assumptions are made about the clinical appropriateness of such detail.

Where the eligibility text appears to include both inclusion and exclusion sections, the text is split accordingly, and each part is analysed separately. The processing relies on straightforward string matching rather than sophisticated semantic parsing, meaning that only well-formatted criteria are reliably separated. From this division, the number of sentences in each section is counted using standard sentence detection tools. These counts are retained as two additional features, capturing the structure of the eligibility logic in a form that can be compared across trials.

In parallel, both the primary and secondary outcome fields are examined. These are expected to contain lists of defined outcome measures, though their presence and quality vary considerably across records.

The total number of primary and secondary outcomes is recorded for each trial, allowing for later examination of whether outcome multiplicity correlates with trial status or sponsor type. A binary flag is also included to indicate whether a trial reports any primary or any secondary outcomes at all, which can be useful when filtering for minimal completeness or assessing reporting discipline.

All of these features are designed to summarise key textual content without introducing opaque representations or complex embeddings. They allow the richness of protocol narratives to be captured in simple, transparent variables that can inform further modelling or descriptive analysis, while ensuring that the original free-text remains accessible for qualitative interpretation if needed.

7. Categorical Encoding and Trial Classification

Several key characteristics of clinical trial design are recorded as categorical variables. These include masking strategies, randomisation methods, intervention models, and sponsor classifications. While such information is invaluable for understanding the structural features of a trial, it is often presented in inconsistent formats, with spelling variation, missing values, or overlapping categories. In order to support reliable analysis, these fields are transformed into well-defined categorical encodings through a set of rule-based procedures.

All relevant categorical fields are first converted into lower-case strings to ensure uniformity in representation. Values such as "Open", "open label", or "Open-Label" are therefore standardised and treated identically. This step avoids the fragmentation of conceptually identical entries across multiple labels. The same approach is applied to randomisation methods, sponsor types, and intervention models.

Sponsor class, in particular, is further refined by assigning each trial to a higher-level sponsor group. These groups are defined as industry, government, or non-industry. Industry sponsors are identified explicitly. Government-affiliated trials are grouped based on known categories, such as the National Institutes of Health or other federal entities. Trials not falling into either of these are placed in the non-industry category, which includes academic, non-profit, and other private sector organisations not classified as industrial.

Interventions are analysed more extensively. Each intervention is examined to determine its type, such as drug, biological, device, behavioral, or procedure. Counts are calculated for each type present within a given trial. These counts are then used to create a set of variables that reflect not only the presence or absence of each type, but also their relative proportions. For example, the proportion of drug-related interventions within a trial is calculated and recorded as a numerical feature.

Beyond individual intervention types, trials are also grouped according to broader intervention patterns. Trials containing only drugs are categorised distinctly from those that contain combinations, such as drug and device, or those using multiple types without any drugs. This grouping enables researchers to study trial characteristics at a higher level of abstraction without losing track of the underlying intervention diversity.

An additional binary representation is created for each core intervention type to indicate whether that type is present in a trial. This results in features such as has_device_intervention or has_biological_intervention, which are useful for simple filtering and summarisation.

The resulting classification system provides a clear, interpretable summary of trial design features, enabling comparisons both within and across sponsor categories, intervention strategies, and structural designs. All encodings are deterministic, reproducible, and grounded in a straightforward set of parsing and grouping rules that align with registry semantics.

8. Condition Grouping via Biomedical Ontology

Clinical trials are commonly annotated with one or more condition labels describing the targeted disease or health state. These labels, while informative, are often drawn from free-text entries and exhibit considerable variation in spelling, phrasing, and granularity. For instance, one trial might list “lung cancer”, another “non-small cell lung carcinoma”, and a third “pulmonary malignancy”. Although these terms refer to related or overlapping conditions, they may be treated as distinct unless harmonised.

To address this issue, the preprocessing workflow incorporates a grouping mechanism grounded in the Medical Subject Headings (MeSH) hierarchy. Specifically, each condition label is normalised by removing punctuation, converting to lower case, and stripping extraneous whitespace. The resulting term is then matched to a curated index of conditions that maps each entry to one or more top-level MeSH disease categories, denoted by their corresponding C-codes.

The matching procedure proceeds through a series of increasingly permissive steps. Initially, it attempts to find an exact match with the controlled condition index. If no match is found, the process continues with fuzzy string matching to identify closely related terms based on similarity scores. Where necessary, the text is further decomposed into individual tokens, each of which is then matched independently. If all other steps fail, a fallback is used that searches for substrings within the index. If none of these methods yield a match, a default label of “OTHER” is assigned.

Each trial may be assigned to more than one top-level group if its listed conditions map to multiple distinct disease categories. The final result is a set of high-level condition groups associated with each trial, which enables stratification by therapeutic area, such as oncology, cardiology, or infectious disease. This grouping allows for more meaningful comparisons between trials and facilitates domain-specific analysis without depending on the inconsistent wording found in raw condition labels.

Importantly, this classification does not discard the original condition entries but rather supplements them with a structured ontology-based mapping. As a result, the dataset retains its clinical richness while gaining the consistency needed for large-scale analytical work.

9. Handling of Missing and Derived Fields

Missingness is a pervasive feature of clinical trial records. Certain variables may be absent due to omission at the point of registration, delayed updates, or inconsistencies in reporting practices. A key goal

of the preprocessing phase is not to conceal or forcibly impute missing values, but to identify them clearly and handle them in a manner that preserves the interpretability and integrity of the data.

One of the most salient derived variables in the dataset is trial duration, defined as the number of days between the reported start date and the reported completion date. When both dates are present and valid, this value is calculated directly and added as a new field. Where either date is missing or invalid, the duration is left blank, and the record is flagged with a Boolean indicator to identify the presence of missing duration data.

In a separate process, records missing duration information are later revisited. A lookup table is constructed from the subset of trials with known durations, keyed by trial identifier. For trials with missing duration but known identifiers, the corresponding value from the lookup is retrieved and assigned. This approach does not involve statistical imputation, but rather leverages known values from parallel sources to fill gaps in an exact, traceable manner.

No other missing fields are altered during preprocessing. Fields that are absent from the source data remain so in the output, ensuring that downstream models or analyses can adopt bespoke missing data strategies appropriate to their assumptions and design. No assumptions are made about why a value is missing, and no effort is made to fabricate or guess likely values.

This conservative handling of missingness ensures that the final dataset remains faithful to the underlying registry content, while offering enough scaffolding to support robust analytical workflows. Derived values are introduced only where logically sound and computationally safe to do so. Where completeness cannot be achieved, the absence is made explicit, rather than disguised.

10. Parallelisation and Performance Considerations

Processing clinical trial data at scale presents practical challenges. Although individual trial records are modest in size, their combined volume across multiple disease areas can be substantial, particularly when the dataset extends to thousands of entries. To address this, the preprocessing pipeline adopts a parallel processing strategy designed to improve throughput without compromising reproducibility.

Trials are processed concurrently using a pool of independent workers, each responsible for executing the full preprocessing logic on a single record. This includes field normalisation, feature extraction, condition grouping, and classification. The parallel execution is orchestrated using a multi-process architecture, which takes advantage of available CPU resources to accelerate the pipeline.

Each trial is handled in isolation, with no dependencies across records. This allows the preprocessing to scale linearly with the number of available workers, subject to system constraints. If a particular record causes an error during processing, the exception is captured and logged without halting the entire pipeline. This fail-safe design ensures that isolated inconsistencies do not derail the overall workflow.

Once all records have been processed in parallel, the results are gathered and passed through a final quality control step. At this stage, columns with a high proportion of missing values are identified and, where appropriate, excluded from the final output. The current implementation removes variables for

which more than two percent of entries are missing, though this threshold may be adjusted depending on the needs of the downstream analysis.

The parallelised approach offers a pragmatic balance between computational efficiency and methodological clarity. It ensures that large volumes of data can be processed in a reasonable time frame, without resorting to opaque batch operations or sacrificing the interpretability of individual trial features.

11. Output Structure and Feature Schema

The culmination of the preprocessing phase is a structured dataset in which each trial is represented by a single, enriched record containing both original registry information and a set of derived features. The aim of this representation is not simply to tidy the data, but to provide a coherent and interpretable foundation upon which subsequent modelling, analysis, or visualisation can reliably proceed.

Each output record preserves the original content of the trial wherever possible. Fields such as trial identifier, masking strategy, intervention type, eligibility text, sponsor name, and study phase are retained in their normalised form. Alongside these original fields, the pipeline appends a set of new variables that capture additional information derived through linguistic analysis, duration calculation, categorical grouping, and biomedical mapping.

The output includes numerical features such as trial duration in days, the number of primary and secondary outcome measures, the token count of eligibility criteria, and sentence counts for inclusion and exclusion sections. It also includes Boolean flags indicating whether key components are present, such as any defined primary outcome or FDA regulatory oversight.

Categorical fields are encoded in a way that facilitates comparison across trials. For example, sponsor types are classified into higher-level categories, while intervention types are grouped into broader archetypes that reflect whether the trial is drug-only, device-only, or a more complex mixture. These fields are represented using clearly labelled variables that allow for grouping, filtering, and aggregation in a transparent manner.

Condition groupings derived from the MeSH ontology are recorded as lists of high-level disease codes. This enables researchers to stratify analyses by therapeutic area without relying on inconsistent free-text labels. Although a trial may belong to more than one condition group, the output format supports this multiplicity without losing information.

All data is stored in a machine-readable JSON format, with each trial occupying one entry in a list. This structure allows for seamless loading into statistical or machine learning environments while retaining sufficient transparency to support manual inspection when needed.

The feature schema produced by the pipeline is not fixed, but rather extensible. It reflects a design philosophy in which every derived field is traceable, interpretable, and grounded in a reproducible logic. This output serves as both a practical resource for analysis and a methodological statement about the value of deliberate, interpretable feature engineering.

12. Limitations and Methodological Considerations

While the preprocessing framework outlined in this report succeeds in transforming a heterogeneous collection of trial records into a structured and analytically coherent dataset, it is important to acknowledge the limitations inherent in both the data and the methods employed. These limitations do not undermine the utility of the output but rather define the boundary within which its use remains appropriate and scientifically defensible.

The most immediate limitation arises from the variability and incompleteness of the source data. Although the pipeline includes logic for detecting and handling missing fields, it does not attempt to statistically impute missing values beyond a direct recovery of trial duration where possible. Records lacking valid start or end dates, for instance, remain flagged as incomplete even if their other features are intact. This choice reflects a cautious approach, prioritising transparency over speculative correction.

A second limitation lies in the simplicity of the text processing techniques. Token and sentence counts are used to characterise eligibility criteria and outcome sections, but no attempt is made to extract clinical entities, detect logical operators, or distinguish between criteria types beyond the basic inclusion and exclusion split. As a result, these features are best understood as structural summaries rather than semantic representations.

There are also limits to the condition grouping logic. While the mapping of condition labels to MeSH-based disease categories greatly improves consistency, it is not immune to failure. Some condition names are too vague, novel, or idiosyncratically expressed to yield confident matches. In such cases, the fallback category “OTHER” is applied. Although the process includes steps for fuzzy matching and partial token comparison, the ontology coverage is necessarily imperfect and cannot capture every nuance of clinical labelling.

Categorical encoding, though useful for standardisation, inevitably involves simplification. For example, the grouping of sponsors into “industry”, “government”, or “non-industry” masks finer distinctions between academic institutions, charities, and cooperative groups. Similarly, intervention archetypes, while helpful for stratification, may overlook clinically meaningful differences within a given class.

Finally, despite the use of parallel processing to improve efficiency, the pipeline remains dependent on system resources and may not scale indefinitely without modification. In the event of malformed records or unexpected field combinations, errors are captured and logged, but there is no mechanism for automated correction or human-in-the-loop inspection during runtime.

Taken together, these limitations reinforce the importance of treating the output not as a perfect representation of trial structure, but as a well-prepared starting point for responsible analysis. It remains the role of the researcher to assess feature relevance, account for uncertainty, and apply suitable interpretive care in concluding the processed data.

13. Conclusion

This report has presented a systematic approach to the preprocessing of clinical trial protocol data, with a focus on transforming unstructured and inconsistently reported registry records into a structured, interpretable, and analysis-ready format. Each step of the pipeline has been developed with clarity, transparency, and methodological caution, ensuring that the resulting dataset preserves the complexity of clinical trial design while rendering it suitable for robust downstream analysis.

Rather than relying on abstract imputation or opaque transformations, the preprocessing strategy is grounded in traceable logic and reproducible methods. Text fields are summarised using straightforward linguistic metrics. Categorical variables are harmonised through explicit mapping rules. Eligibility criteria and intervention strategies are classified in a way that respects both registry semantics and biomedical structure, without overreaching the limits of the source data.

Particular care has been taken to address missingness, manage variable formats, and identify repeated records. The use of controlled vocabularies, such as MeSH-based disease groupings, further enhances the interpretability of condition labels, allowing for higher-level comparison across therapeutic areas. The pipeline is also computationally efficient, employing parallelisation to handle large datasets without compromising data integrity.

Yet, as with any structured extraction effort, there remain caveats and constraints. The underlying data is imperfect. Some labels defy categorisation, some fields are sparsely populated, and not every design decision made by trial sponsors is recoverable through metadata alone. What this preprocessing offers is not completeness, but coherence, a way to make sense of what has been recorded, and a foundation for asking more informed, tractable questions about the trials that shape clinical knowledge.

In viewing preprocessing as an intellectual and scientific task in its own right, rather than a mere technical step, this work highlights the central role of preparation in any data-centric investigation. Before one can model, one must listen carefully to the structure of the data. This report represents that listening is deliberate, patient, and grounded in the understanding that no reliable conclusion can emerge without first attending to the form in which information arrives.