TrialIQ Data Extraction Report: Structured Metadata Retrieval for Predictive Clinical Trial Design

Author: Francis Osei
Email: francis@bayezian.com

1. Introduction

Clinical trials are the bedrock of evidence-based medicine, yet their execution is often beset by inefficiencies, premature terminations, and failure to meet primary endpoints. A substantial proportion of interventional trials, particularly those in early- to mid-phase development, do not progress to completion or yield conclusive results. While the prevailing narrative frequently attributes such outcomes to therapeutic inefficacy, emerging evidence suggests that design-related factors, including complex eligibility criteria, inadequate power, and operational burden, play a significant role in undermining trial success.

Despite growing access to structured protocol metadata through registries such as ClinicalTrials.gov, the analytical exploitation of this data remains surprisingly limited. Most assessments of trial failure are retrospective, offering post hoc rationales rather than prospective tools for mitigation. The underutilisation of protocol-level design features available before participant enrolment represents a critical gap in the methodological landscape of clinical development.

There is, therefore, a pressing need to harness this metadata more systematically to identify early signals of failure risk. By interrogating trial design features at scale, encompassing aspects such as study phase, masking strategies, eligibility formulations, and sponsor profiles, it may be possible to anticipate the likelihood of trial disruption or futility before resources are committed.

This research seeks to contribute to that emerging paradigm by developing a reproducible approach for extracting and structuring protocol metadata across a broad spectrum of disease domains. In doing so, it aims to lay the groundwork for future empirical models capable of quantifying design-related risk and informing more resilient trial planning.

2. Objectives

The central aim of this research is to enable the early identification of clinical trial designs that may be predisposed to failure, using protocol metadata available before enrolment. Rather than relying on trial outcomes or interim analyses, the approach focuses on information captured at registration, offering a pre-emptive, design-centric perspective on trial risk.

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

1. To curate a high-quality dataset of clinical trial protocols spanning multiple therapeutic areas, using disease-specific search criteria derived from biomedical ontologies.
2. To extract structured variables from protocol-level metadata, including key dimensions of trial design, recruitment criteria, sponsor characteristics, and regulatory oversight.

3. To filter and retain trials of analytical relevance, applying inclusion criteria that reflect methodological robustness and temporal recency.
4. To produce a harmonised, analysis-ready dataset that preserves the heterogeneity of trial design choices while facilitating subsequent statistical or machine learning modelling.

While this work does not seek to model trial success directly, it establishes a methodological pipeline for generating the inputs required to do so. In doing so, it provides a foundation for future efforts aimed at predicting trial failure, benchmarking design complexity, or supporting adaptive protocol development.

3. Data Source and Scope

The primary data source for this research is *ClinicalTrials.gov*, a public registry maintained by the U.S. National Library of Medicine. The registry provides structured protocol information for interventional and observational studies across a wide range of therapeutic areas and geographical regions. Each record includes detailed metadata describing the study design, eligibility criteria, sponsor involvement, outcome measures, and regulatory oversight.

To ensure analytical consistency and clinical relevance, this study focuses exclusively on interventional trials associated with human diseases, as defined within the Medical Subject Headings (MeSH) hierarchy. Specifically, descriptors belonging to category "C" of the MeSH tree denoting disease entities were used as search terms to retrieve relevant trials. This ontological filter ensures that the dataset captures biologically grounded, therapeutically relevant studies, rather than exploratory or non-clinical investigations.

The scope of trial inclusion is further constrained by three methodological considerations:

1. Temporal relevance: Only trials with a registered start date within the past ten years are considered, in order to capture contemporary design practices.
2. Design clarity: Trials must report a valid study phase and belong to one of the accepted completion statuses, namely, *Completed*, *Terminated*, or *Withdrawn*. This ensures that trials are either finalised or formally discontinued, avoiding ambiguous or incomplete entries.
3. Protocol completeness: Records lacking essential metadata such as outcome definitions, eligibility criteria, or intervention descriptions are excluded from analysis.

By applying these filters, the resulting dataset comprises a curated subset of trials whose design features are sufficiently complete and standardised to permit downstream comparative analysis. Importantly, this selection strategy prioritises data quality over volume, reflecting a preference for interpretability and methodological transparency.

4. Variable Extraction and Design Features

To enable comparative analysis of trial designs, a set of structured variables was systematically extracted from each eligible protocol record. These variables were selected to reflect critical dimensions of study architecture, operational planning, and regulatory positioning, all of which are hypothesised to influence trial success or early termination.

The extracted metadata is grouped thematically as follows:

4.1 Study Identification and Classification

- Title: A brief textual description of the trial’s objective.
- NCT Identifier: The unique trial registration number.
- Study Type: Classification as interventional, observational, or other.
- Study Phase: Developmental stage (e.g., Phase I, II, III, or IV).

4.2 Design Complexity

- Masking Strategy: Whether the trial uses blinding (e.g., single, double, or open label).
- Randomisation: Presence and nature of participant allocation methods.
- Intervention Model: Parallel, crossover, or factorial assignment structures.
- Primary Purpose: Therapeutic intent, including treatment, prevention, or diagnostic exploration.
- Number of Arms: Count of distinct intervention or comparator groups.

4.3 Eligibility and Recruitment

- Eligibility Criteria: The full inclusion and exclusion criteria are provided in free-text form.
- Minimum and Maximum Age: Reported age thresholds for participation.
- Sex: Whether the study targets males, females, or all sexes.
- Healthy Volunteers: Whether participants are healthy individuals.

4.4 Interventions and Outcomes

- Intervention Types: Classifications (e.g., drug, procedure, behavioural) and names of interventions.
- Primary Outcomes: Key efficacy or safety endpoints as specified at registration.
- Secondary Outcomes: Additional exploratory or supporting measures.

4.5 Sponsorship and Oversight

- Sponsor Name and Class: Lead sponsor identity and affiliation type (e.g., academic or industry).
- Collaborating Institutions: List and count of non-lead collaborators.
- Site Count: Number of study locations.
- Regulatory Oversight: Indicators of FDA regulation for drugs or devices; presence of a Data Monitoring Committee (DMC); and listing of oversight authorities.

4.6 Temporal Markers and Access

- Start and Completion Dates: Structured date formats for trial initiation and conclusion.
- Actual Enrollment: Number of participants enrolled if reported as final.
- Expanded Access: Indication of whether the trial offers access to the investigational treatment outside the context of the trial.

By consolidating these variables into a harmonised schema, the dataset captures both the structural and operational contours of trial design. This provides a rich empirical basis for examining how different configurations in terms of eligibility strictness, sponsorship, or design strategy may correlate with trial outcomes in future analytical studies.

5. Inclusion Criteria and Data Filtering

A rigorous filtering process was applied to ensure that only trials of sufficient quality, completeness, and analytical relevance were retained for further analysis. Given the heterogeneity of trial records within ClinicalTrials.gov, this step was essential to mitigate noise, enhance comparability, and preserve the interpretive validity of downstream inferences.

The inclusion criteria were designed around three principal dimensions:

5.1 Temporal Relevance

Trials were restricted to those with a recorded start date within the past ten years. This decision reflects an intention to capture current trial design practices, regulatory expectations, and methodological conventions. Older studies, while potentially valuable in historical contexts, were excluded to reduce the risk of incorporating obsolete or incomparable design elements.

5.2 Completion Status

Only trials with a clear and finalised overall status were retained. Specifically, eligible records were those marked as:

- *Completed*: The trial concluded as planned and reached its defined endpoint(s).
- *Terminated*: The trial was discontinued before completion, typically due to safety, efficacy, or recruitment concerns.
- *Withdrawn*: The trial was halted before participant enrolment began.

Excluding ongoing, suspended, or ambiguous trials allowed for a more definitive classification of design features relative to trial fate, which is critical for any future development of predictive models.

5.3 Methodological Completeness

Trial records were further filtered based on the presence of core protocol components. Trials lacking any of the following were excluded:

- Stated study phase (to allow stratification by developmental stage)
- Defined intervention(s) (to ensure structural comparability)
- Reported outcome measures (primary at minimum)
- Registered start date in a valid format

Additionally, trials listing their phase as "N/A" were excluded, as this designation typically indicates early feasibility or non-interventional designs not amenable to standard success/failure classifications.

Together, these criteria yielded a refined dataset of trials that are not only methodologically sound but also sufficiently detailed to support comparative analysis. This filtration process ensures that the integrity of the resulting metadata is upheld, providing a solid empirical foundation for evaluating how specific design characteristics may relate to trial outcomes.

6. Data Structuring and Output Format

Following extraction and filtering, the protocol-level metadata was consolidated into a structured, machine-readable format to facilitate subsequent quantitative analysis. Particular attention was paid to ensuring internal consistency across variables, the preservation of original registry semantics, and the usability of the resulting dataset for future modelling tasks.

6.1 Record Structure

Each trial entry was encoded as a structured data object comprising a fixed set of fields corresponding to the thematic domains outlined in Section 4. These include identifiers, design specifications, intervention details, eligibility criteria, outcomes, sponsorship information, regulatory oversight markers, and temporal variables. Where appropriate, nested structures were flattened for ease of tabular representation, while lists (e.g., outcome measures, collaborators) were preserved in serialised form.

6.2 Handling of Textual Content

Free-text fields such as eligibility criteria and summaries were retained in full, enabling both human interpretability and potential use in natural language processing applications. No semantic compression or keyword reduction was applied at this stage to preserve the richness of the source material and allow for downstream flexibility in feature engineering.

6.3 Categorical Standardisation

Categorical fields such as masking strategies, intervention models, sponsor types, and regulatory labels were standardised according to registry documentation. In cases of known variation in spelling or format (e.g., “Double” vs. “Double-blind”), values were harmonised to a canonical form to reduce redundancy and enhance grouping fidelity during analysis.

6.4 Missing Data and Null Representation

Where individual trials lacked specific metadata (e.g., a missing maximum age or unreported enrolment figure), missingness was explicitly encoded as null values rather than inferred or imputed. This design choice was made to ensure transparency and permit researchers to apply bespoke missing data strategies suitable to their intended analytic frameworks.

6.5 Output Format

The final dataset was exported in JSON format, offering flexibility for integration into a wide range of statistical, machine learning, or visualisation environments. Each entry corresponds to a single clinical trial and encapsulates all retained fields following extraction and cleaning. The format supports both row-wise inspection and aggregate transformation, making it well-suited for exploratory and predictive workflows alike.

By structuring the data in this way, the dataset provides a coherent, analysis-ready foundation for investigating how protocol design characteristics may correlate with operational success, trial termination, or broader measures of research efficiency.

7. Limitations and Methodological Considerations

While the dataset and methodology presented offer a robust foundation for exploring the relationship between clinical trial design and trial outcomes, several limitations must be acknowledged. These caveats pertain to both the nature of the source data and the structure of the extraction process.

7.1 Redundancy and Repetition

One notable limitation is the presence of duplicate or near-duplicate trial entries associated with different search terms. Owing to the overlap of medical conditions across disease taxonomies, particularly in cases where trials are tagged with multiple MeSH descriptors, some records appear multiple times across different retrieval cycles. Although each entry is structurally identical, the repetition inflates the raw volume of extracted trials and necessitates deduplication before any aggregate analysis. This redundancy is an artefact of the registry's indexing logic and underscores the importance of careful post-processing.

7.2 Registry-Level Inconsistencies

ClinicalTrials.gov, while comprehensive, exhibits variable completeness across records. Not all trials report the same depth of detail, particularly concerning eligibility criteria, oversight information, or sponsor roles. This heterogeneity limits the comparability of certain features across the dataset and may introduce bias if analyses are performed on subsets with systematically more complete metadata.

7.3 Designation of Trial Status

Although the filtering procedure excludes ambiguous or ongoing studies, it remains reliant on sponsor-submitted updates to determine a trial's true status. Trials labelled as “Terminated” or “Withdrawn” may do so for reasons unrelated to design quality, such as strategic reprioritisation, funding withdrawal, or shifts in regulatory guidance, which may not be discernible from the metadata alone.

7.4 Absence of Outcome Labels

Importantly, this study does not incorporate direct measures of trial success or endpoint achievement. The dataset reflects protocol design characteristics alone, without linking them to efficacy readouts or publication status. As such, it should be regarded as a design-focused resource, suitable for structure-outcome inference only when paired with external success indicators in future research.

7.5 Language and Ontology Constraints

By relying exclusively on MeSH disease terms in English, the dataset may under-represent trials indexed using non-standard nomenclature or condition terms outside the MeSH vocabulary. While this enhances conceptual coherence, it may omit relevant studies from emerging or rare disease areas.

Taken together, these limitations do not undermine the utility of the dataset, but rather inform its responsible use. Any predictive or comparative modelling built upon this resource should account for these constraints, and wherever possible, incorporate mechanisms for validation against independent outcome indicators.

8. Conclusion

This study presents a methodologically grounded approach to curating structured protocol metadata from registered clinical trials, with a particular emphasis on extracting variables that capture the complexity and architecture of trial design. By leveraging a disease-informed search strategy and applying systematic inclusion filters, the resulting dataset offers a high-resolution snapshot of design decisions across a diverse portfolio of interventional studies.

The work addresses a notable gap in the current research landscape, namely, the underutilisation of pre-enrolment protocol features for early risk assessment. While most analyses of clinical trial performance are retrospective and outcome-driven, this research contributes a forward-facing foundation that enables empirical investigation into how design characteristics may influence trial viability.

The dataset itself is comprehensive, interpretable, and well-suited for integration into predictive modelling workflows or design benchmarking tools. Although several limitations remain, including data repetition, occasional missing fields, and the absence of explicit outcome labels, the approach provides a reproducible and extensible basis for future methodological work.

Ultimately, this research underscores the analytical value of treating trial protocols not merely as administrative documents, but as data-rich artefacts capable of informing smarter, more resilient clinical development strategies. By doing so, it opens a path towards more efficient evidence generation, better resource allocation, and reduced rates of trial failure in the long term.