TRIALIQ Descriptive Statistical Characterisation of Structured Clinical Trial Metadata: A Comparative Summary of Full, Training and Test Sets

Author: Francis Osei
Email: Francis@bayezian.com

1. Introduction and Rationale

Before any modelling can be responsibly undertaken, it is essential to understand the structure, balance, and distribution of the data upon which those models will rest. This is especially true in the context of clinical trial metadata, where design choices, regulatory markers, and trial complexity vary widely, not just in content, but in how that content is reported. The value of this summary, then, lies not in abstraction, but in clarity: it offers a grounded picture of what is present, what is missing, and what patterns quietly emerge when thousands of trials are read side by side.

This report presents a descriptive statistical characterisation of a processed dataset of interventional clinical trials. The data, having undergone careful preprocessing and normalisation, is now summarised in full and also examined in its training and test subsets. The aim is not to anticipate model results, but to inform their design by reflecting on the shape and content of the data they will ingest.

Descriptive statistics are more than a formality. They help reveal where age eligibility is concentrated, how often trials report their outcomes properly, and whether masking and randomisation are common practice or notable exceptions. They expose sparsity, imbalance, and redundancy, and equally, they confirm where the dataset reflects the clinical reality it seeks to describe.

By analysing both continuous and categorical features, including trial durations, intervention types, sponsor affiliations, and condition groupings, this report invites the reader into a thoughtful exploration of the dataset's texture. Special attention is paid to how these features are distributed across the training and test sets, ensuring that any evaluation carried out later can be understood in the context of what the model has and has not seen.

In short, this section sets the scene. It gives us not just numbers, but a way to read the dataset with care and to proceed, not just with computation, but with context.

2. Overview of the Analytical Dataset

The analytical dataset consists of structured records derived from interventional clinical trials registered on ClinicalTrials.gov. Each trial entry includes protocol-level metadata that has been preprocessed to ensure consistency in format, completeness in core variables, and interpretability for statistical and machine learning applications.

Following preprocessing, the dataset was filtered to include only trials with a calculable duration. This filtering step ensures that time-based variables are well defined and that records used in modelling or descriptive analysis contain sufficient information for comparative evaluation.

The resulting dataset was partitioned into a training set and a test set using an 80:20 stratified random split. This partitioning was carried out at the record level, preserving the statistical properties of the

original dataset in each subset. The training set comprises 32,541 trials, while the test set includes 8,136 trials. These partitions are mutually exclusive and collectively exhaustive concerning the filtered dataset.

Each trial record contains a standard set of variables, including trial phase, masking method, randomisation category, primary purpose, intervention model, sponsor classification, regulatory oversight status, and structured demographic eligibility. In addition to these, the dataset includes derived features such as trial duration in days, token and sentence counts from eligibility criteria, and intervention-based indicators, including the drug intervention ratio and overall intervention count.

Condition annotations associated with each trial were mapped to top-level MeSH categories, allowing each record to be associated with one or more high-level disease domains. The number of condition groups per trial is explicitly recorded to support analysis of therapeutic scope and disease complexity.

All records in both the training and test partitions contain complete data for the variables described in this report. This ensures that descriptive summaries and downstream models can be constructed without the need for additional imputation or filtering.

The summaries presented in the following sections are computed independently for the full dataset, the training subset, and the test subset. This comparative approach allows for detailed inspection of dataset properties and verification of balance between the modelling partitions.

3. Summary of Continuous Variables

This section presents descriptive statistics for all continuous variables in the dataset. These include numeric measures relating to trial structure, eligibility criteria, intervention planning, and study duration. The statistics are reported separately for the full dataset, the training partition, and the test partition. Each variable is summarised using standard measures: count, mean, standard deviation, minimum, first quartile, median, third quartile, maximum, and count of missing values.

The number of study arms per trial (arm_count) exhibits a right-skewed distribution. Across the full dataset, the median number of arms is two, with the majority of trials falling between one and three arms. A small subset of trials includes highly complex designs with more than ten arms.

Minimum age (minimum_age) is consistent across the dataset, with the majority of trials setting the lower age threshold at 18 years. The median and lower quartile values confirm this default threshold, while a minority of paediatric trials include younger participants. Maximum age (maximum_age) varies more widely, with values ranging from early childhood to over 300 years, indicating either open-ended eligibility or inclusion of aggregated elderly populations. Most trials cluster around 85 to 100 years as a standard upper limit.

Trial duration, measured in days (trial_duration_days), shows substantial variability. The median duration is approximately 608 days in the full set, with an interquartile range spanning from roughly 300 to 1080 days. Extremely long trials are present, with durations exceeding 6,000 days in a few cases. No durations are missing, as only complete records were retained for analysis.

Text-derived features include the number of tokens in the eligibility criteria field (eligibility_token_count) and the number of sentences in the inclusion and exclusion sections (inclusion_criteria_count, exclusion_criteria_count). These variables provide a measure of protocol complexity. The average eligibility description contains approximately 466 tokens, with notable variation. Sentence counts for inclusion and exclusion criteria are similarly variable, with medians of eight and ten sentences, respectively.

The number of reported outcome measures is recorded separately for primary (primary_outcome_count) and secondary (secondary_outcome_count) outcomes. Most trials list one or two primary outcomes, with a small number including unusually large sets, in some cases exceeding one hundred. Secondary outcomes are more numerous on average, reflecting exploratory endpoints and safety monitoring requirements.

The ratio of drug interventions to total interventions (intervention_type_drug_ratio) is included as a numeric proportion ranging from 0 to 1. Trials with exclusively drug-based interventions have a ratio of one, while those incorporating devices, behavioural strategies, or procedures register lower values. The mean ratio is approximately 0.76, indicating that most trials involve at least one pharmacologic component.

No continuous variable in the dataset contains missing values at the time of summarisation. All entries retained in the training and test sets are complete concerning the fields reported here.

4. Distribution of Categorical Attributes

This section presents the frequency distributions of key categorical variables within the dataset. These variables represent discrete descriptors of trial design, sponsor affiliation, regulatory oversight, and methodological intent. Each distribution is summarised across the full dataset, the training set, and the test set. All values have been normalised to account for potential spelling inconsistencies and unstructured input formats during preprocessing.

The trial phase variable is dominated by entries categorised as Phase I, Phase II, and Phase III. Early Phase I and Phase IV are also present, though in smaller proportions. The distribution of phases is consistent across the training and test sets, reflecting uniform representation of trials across development stages.

Masking methods are diverse, with the largest proportion of records classified as unmasked or labelled as "other", indicating either no masking or ambiguous reporting. Quadruple and double masking strategies are commonly used in well-controlled studies, while single and triple masking appear less frequently. This distribution indicates that a significant number of trials either do not employ masking or do not specify their masking protocol.

The randomisation category variable reflects whether trials use random allocation, do not randomise, or report no allocation method. Randomised trials form the majority, while a sizeable portion of records either omit randomisation information or specify that randomisation is not applicable. This suggests variation in methodological rigour and reporting practices across the dataset.

The primary purpose is dominated by treatment-focused studies. Prevention, diagnostic, and basic science trials are also represented, albeit in smaller numbers. Other less frequent categories include supportive care, health services research, and screening. This distribution confirms the central focus on therapeutic intervention in the registered dataset.

Sponsor class is stratified into industry, government, and non-industry categories. Industry sponsors account for approximately half of all records. Government sponsors, including federal bodies and public research institutions, represent a much smaller proportion. The remaining entries are grouped under non-industry classifications, which include academic and independent institutions.

Regulatory fields include binary indicators for FDA-regulated drug and device trials, the presence of a data monitoring committee, and the availability of expanded access programmes. Most trials report no expanded access and either confirm or deny FDA oversight, though a subset of entries lacks definitive regulatory status and is categorised as "other". The proportion of trials with oversight by a data monitoring committee is moderate, suggesting variable implementation of independent trial governance.

The intervention model variable includes categories such as parallel assignment, single group assignment, crossover, factorial, and sequential designs. Parallel assignment is the most common model, reflecting standard two-arm trial structures. Single-group designs are frequently used in early-phase or feasibility trials. Crossover and factorial designs are relatively rare.

All categorical distributions are consistent between the training and test partitions, with no evidence of significant imbalance. This supports the suitability of the current partitioning scheme for modelling and evaluation purposes.

5. Condition Group Prevalence Across the Corpus

Each trial in the dataset is annotated with one or more condition labels, reflecting the disease areas targeted by the intervention. These labels have been mapped to top-level categories defined by the Medical Subject Headings (MeSH) hierarchy, enabling consistent grouping of trials by therapeutic domain. The result is a set of binary indicators for each MeSH disease category, denoted by their corresponding C-codes.

The most prevalent condition group is C23, which corresponds to pathological conditions and signs. This group appears in over 28 percent of all records and spans a wide range of symptom-driven or syndromic presentations. It is followed by C01 (bacterial infections and mycoses), C04 (neoplasms), and C10 (nervous system diseases). These four categories collectively cover a significant proportion of the dataset, reflecting the clinical and research emphasis placed on infectious disease, oncology, and neurological disorders.

Other commonly observed groups include C08 (respiratory tract diseases), C06 (digestive system diseases), and C17 (skin and connective tissue disorders). Cardiovascular diseases (C14), immune system diseases (C20), and musculoskeletal conditions (C05) are also well represented. Less common categories, such as C24 (occupational diseases), C21 (animal diseases), and C22 (pathologic conditions, signs and symptoms), appear infrequently and contribute minimally to overall representation.

Each trial may be associated with multiple condition groups. The distribution of condition group count indicates that the majority of trials reference one or two high-level disease categories. A smaller number of trials are associated with three or more categories, typically reflecting comorbid populations or broad intervention scopes. In the test set, approximately 45 percent of records reference a single disease group, while roughly 31 percent reference two, and the remainder span a wider array of groupings.

The distributions of condition groups are stable across the training and test sets. This consistency suggests that the random partitioning procedure preserved the therapeutic diversity of the original dataset and did not introduce domain-specific biases.

These groupings are critical for stratified analysis, as they allow researchers to evaluate model performance across distinct clinical domains and to detect whether certain trial design patterns are overrepresented in particular disease areas. They also serve as a foundation for investigating trial success rates, duration trends, and sponsor types by therapeutic focus.

6. Comparative Observations Between Train and Test Sets

The training and test partitions were constructed using a stratified random sampling procedure applied to the processed dataset. The purpose of this division is to support the development and evaluation of predictive models while preserving the underlying distributions of key clinical and design features. This section presents a comparative analysis of these two subsets, focusing on their alignment across both continuous and categorical variables.

Concerning continuous features, all variables show near-identical distributions between the two partitions. Trial duration, eligibility token count, and intervention type ratios, for example, exhibit equivalent means, standard deviations, and interquartile ranges. The median trial duration is 608 days in both subsets, and the eligibility criteria token count hovers around 466 in each, with similar spread and skewness. These consistencies confirm the absence of systematic sampling bias.

In terms of categorical attributes, proportions remain stable between training and test sets. The phase distribution is consistent across both groups, with Phase II and Phase I trials forming the largest share, followed by Phase III and Phase IV. The masking variable reflects similar profiles in both partitions, with a majority of trials either reporting no masking or categorised under "other". The share of trials using double or quadruple masking remains within one percentage point across partitions.

Randomisation, primary purpose, and sponsor classification also remain aligned. Approximately 64 percent of trials in both partitions are randomised, and treatment remains the dominant primary purpose at over 76 percent. Industry sponsorship accounts for just under 50 percent of all trials in each set. The presence of regulatory oversight (e.g., FDA regulation or data monitoring committees) and expanded access availability also mirrors closely between the two subsets.

Condition group representation, as previously discussed, is preserved across the train and test sets. The top ten most frequent disease groups appear with proportionally equivalent prevalence, and the distribution of condition group counts per trial shows similar patterns.

No meaningful deviations were identified in the distribution of any single variable. There is no evidence of domain concentration, design bias, or regulator-specific imbalance that would require adjustment or resampling. The partitions are therefore suitable for standard model training and evaluation workflows, including domain-agnostic and domain-stratified approaches.

This statistical equivalence supports the use of the training set for general model development and the test set for fair performance evaluation without the need for domain-level normalisation or weighting.

7. Missingness and Coverage Assessment

A key consideration in the preparation and analysis of clinical trial data is the extent and nature of missing values. Incomplete data can compromise statistical inference, limit model generalisability, and introduce structural bias if not addressed systematically. This section evaluates the completeness of the variables retained in the processed dataset, with a focus on assessing whether any patterns of missingness persist within the training or test partitions.

The summary statistics confirm that all continuous variables included in the final dataset are fully populated. This includes trial duration, age ranges, token counts, outcome counts, and numerical ratios. The preprocessing pipeline excluded records without valid trial start and completion dates, thereby ensuring that trial duration could be calculated without imputation. Minimum and maximum age values are also complete, having been normalised and defaulted where necessary during preprocessing.

Categorical variables used in downstream modelling are similarly complete. This includes the trial phase, masking category, randomisation strategy, intervention model, sponsor classification, and regulatory flags. During preprocessing, all free-text or semi-structured categorical fields were standardised and assigned explicit values, including a default label ("other") for entries that could not be matched to a predefined category. This approach avoids the presence of null entries while preserving information about ambiguity or incomplete reporting.

Condition groupings, derived from MeSH mappings, are also available for every trial. Although some trials are assigned only to a broad “other” group due to non-specific or unrecognised labels, the assignment is always present. The number of condition groups per trial is explicitly recorded and complete across the dataset.

Boolean indicators such as healthy volunteer inclusion, FDA regulation, and the presence of a data monitoring committee are consistently represented. Any values that were originally undefined or inconsistently reported have been standardised into a tri-state encoding: true, false, or none. This allows users to distinguish between absence, presence, and non-response without resorting to deletion or coercion.

As a result of this rigorous handling of missingness during preprocessing, no variable in the final dataset requires additional filtering or imputation at the point of statistical analysis. All features used in the summary tables are complete across the full, training, and test datasets.

This high level of coverage provides a stable foundation for modelling and exploratory analysis. It ensures that conclusions drawn from the dataset are not distorted by unrecorded values or inconsistently reported trial characteristics.

8. Interpretation of Trial Design Patterns

The dataset presents a clear and structured view of contemporary clinical trial design practices across a diverse range of therapeutic areas. Through the aggregation of descriptive statistics, several design patterns emerge that are consistent across the full, training, and test partitions.

The majority of trials in the dataset are concentrated in Phase I and Phase II, with a gradual tapering in representation as trials progress to later phases. This distribution reflects the typical attrition pattern in drug and device development, where a larger volume of early-phase trials explore safety, dosing, and feasibility, while only a subset advance to confirmatory stages.

Parallel assignment is the dominant intervention model, present in over half of all trials. This design is widely adopted in comparative efficacy and safety studies, particularly those involving pharmacologic agents. Single-group assignment is also prevalent, especially in early-phase and feasibility trials. Crossover, factorial, and sequential designs are less common and tend to appear in more specialised study contexts.

Masking is inconsistently applied. A substantial proportion of trials either do not report a masking strategy or are classified under a generic “other” label. Among those that do report masking, double and quadruple masking are the most frequently used, particularly in Phase III studies where bias reduction is critical. Single and triple masking are less common, and their use varies by sponsor class and therapeutic area.

Randomisation is implemented in approximately two-thirds of all trials. Non-randomised studies, while still present, are more frequent in early-phase trials and exploratory designs. The remaining trials either do not report their allocation method or indicate that randomisation is not applicable.

The primary purpose is overwhelmingly skewed toward treatment, accounting for over three-quarters of all trials. Preventive and diagnostic studies make up a smaller proportion, with the remainder distributed across basic science, supportive care, and health services research. This distribution aligns with the dominant focus on therapeutic intervention within industry-sponsored and regulatory-driven studies.

Industry sponsorship accounts for roughly half of the dataset, consistent across partitions. Government sponsors, including federal agencies and public research institutions, contribute a much smaller share. The remainder of the trials are associated with non-industry sponsors, including academic groups, hospitals, and research consortia. These patterns are important to track, as sponsor class often correlates with trial phase, resource availability, and methodological rigor.

Intervention type distribution confirms the strong prevalence of drug-based studies. The majority of trials include at least one pharmacologic component, with the average drug intervention ratio approaching 0.76.

Device, behavioural, and procedural interventions are less common and more variable in structure. Mixed intervention models, though present, constitute a minority of the sample.

Taken together, these observations highlight the structural concentration of the dataset around standardised, early-to-mid-phase treatment trials, predominantly driven by pharmacologic research and led by industry sponsors. While the dataset includes methodological variation and therapeutic breadth, it reflects established patterns in clinical development where regulatory requirements, funding structures, and therapeutic priorities shape trial architecture.

9. Conclusion and Relevance to Downstream Analysis

This report has presented a structured statistical characterisation of a curated dataset of interventional clinical trials, following preprocessing and partitioning into training and test sets. The analysis has focused on understanding the distribution of both continuous and categorical variables, the representation of disease domains, and the structural patterns embedded within trial design features.

The dataset is complete across all retained variables, with no missing values in the final set used for modelling. Key dimensions of trial structure, including masking, randomisation, intervention model, sponsor classification, and regulatory oversight, have been standardised and summarised. Condition groups have been mapped to top-level MeSH codes, providing a coherent framework for domain-specific evaluation. The dataset also includes derived features reflecting eligibility complexity and outcome structure.

The training and test partitions are statistically consistent, preserving the distributions observed in the full dataset. No substantial imbalances were detected, and no stratification or domain weighting is required to support valid downstream evaluation. This consistency strengthens confidence in the fairness and reliability of modelling results that will be derived from this data.

The descriptive statistics confirm that the dataset is representative of common clinical trial practices, particularly in early- and mid-phase drug development. It contains a diverse mix of sponsor types, trial phases, and intervention categories, while remaining sufficiently structured to support advanced analytical workflows.

The statistical properties outlined in this report form the empirical foundation for model design, feature selection, and interpretive framing. They support the development of predictive algorithms, risk scoring models, and trial optimisation tools by providing clarity on input distributions and highlighting key design trends.

As a reference document, this summary provides the necessary context for understanding model behaviour and performance. It ensures that analytical decisions are made with full awareness of the underlying data structure and that results are interpreted for the design choices that shape the clinical trial landscape.

Appendix A: Full Summary Table (Aggregated)

This appendix provides the full descriptive summary of all variables computed over the complete processed dataset of 40,677 clinical trials. Continuous variables are summarised using count, mean, standard deviation, minimum, quartiles, and maximum. Categorical variables are reported by absolute frequency and percentage of the total.

The statistics presented here serve as the reference baseline for all comparisons within the training and test partitions. The structure below reflects the original variable types and their statistical treatment in the preprocessing pipeline.

A.1 Continuous Variables (Full Dataset)

Variable	Count	Mean	Std Dev	Min	25%	Median	75%	Max
Arm count	40,677	2.23	1.48	1	1	2	3	56
Minimum age (years)	40,677	18.74	9.55	0	18	18	18	80
Maximum age (years)	40,677	78.42	26.27	0.04	65	85	100	365
Trial duration (days)	40,677	763.20	607.48	0	301	608	1080	6819
Eligibility token count	40,677	465.67	416.49	3	159	331	644	3500
Inclusion criteria count	40,677	12.89	14.17	0	5	8	15	218
Exclusion criteria count	40,677	14.17	13.81	0	5	10	20	135
Primary outcome count	40,677	2.59	4.59	1	1	1	2	214
Secondary outcome count	40,677	6.37	8.94	0	1	4	8	324
Drug intervention ratio	40,677	0.76	0.38	0	0.5	1	1	1

A.2 Selected Categorical Variables (Full Dataset)

Trial Phase

	Phase I	Phase II	Phase III	Phase IV	Early Phase I
Count	13,488	12,478	6,993	6,245	1,473
Percentage	33.16%	30.68%	17.19%	15.35%	3.62%

Masking Category

	Other	Quadruple	Double	Triple	Single
Count	21,341	7,259	5,752	3,780	2,545
Percentage	52.46%	17.85%	14.14%	9.29%	6.26%

Primary Purpose

	Treatment	Prevention	Basic Science	Diagnostic	Supportive Care	Health Services Research	Screening	Device Feasibility	Other
Count	31,014	3,855	1,947	805	774	182	88	44	3
Percentage	76.24%	9.48%	4.79 %	1.98%	1.90%	0.45%	0.22%	0.11%	0.01 %

Condition group

Category	Count	Percent
C23	11442	28.13
C01	8095	19.9
C04	7128	17.52
OTHER	6495	15.97
C10	5320	13.08
C08	4887	12.01
C17	4014	9.87
C20	3783	9.3
C06	3700	9.1
C14	3314	8.15
C12	3192	7.85
C18	3172	7.8

C15	3004	7.39
C19	2481	6.1
C16	2282	5.61
C11	1812	4.45
C05	1747	4.29
C26	969	2.38
C25	963	2.37
C07	924	2.27
C09	555	1.36
C22	240	0.59
C24	20	0.05
C21	7	0.02