Project Title

Quality tolerance limit and duplicated patient investigation in clinical trials

Narrative Summary

The upcoming research project at Cyntegrity focuses on improving the quality control of clinical trials, which are essential for developing new medical treatments. We want to make sure that the data collected in these trials is reliable and accurate. Our study has three main goals:

1. Quality Tolerance Limits (following : ICH E6 R2 5.0.4. Risk control , 5,.0.7 Risk reporting; ICH E6 R3 (draft) 3.10.1.6. Risk Reporting) Threshold (Mean, Upper, Lower limits) Benchmarking: We will carefully examine and compare quality tolerance limits (indicators crucial for assessment of clinical studies risks and/or execution) across different therapeutic areas, phases of clinical trials, and types of trials. By understanding these limits better, we aim to improve the methods used to control and ensure the quality of clinical trial data.

   Cyntegrity is going to investigate QTLs below:

Table 1.

QTLs	Why?	Highest Thresholds Assumption	To check
% or number of subjects who do not meet inclusion/exclusion criteria	To detect whether there may be issues with the screening process and participants being enrolled when they should not have been. Suggests there may be issues with protocol adherence at the site. At study level, it can detect the risk of sites finding it difficult to apply the inclusion/exclusion criteria and the risk to the ability to draw a useful conclusion to the trial.	>=10%	Check the assumption, check for different Therapeutic Areas, understand distribution
% or number of subjects with withdrawal of informed consent	Dropouts due to participants withdrawing informed consent are a subset of all Dropouts from the trial. A higher than expected level of Dropouts for this reason might indicate challenges with the administration practice of the investigational product and/or the side effects (whether or not these are reported as AEs).	>=5%	Check the assumption, check for different Therapeutic Areas, understand distribution
% or number of protocols with incomplete or missing endpoint data	Outcomes/Endpoints are the most important data in a study besides the safety data. They usually are the primary efficacy data (or important secondary efficacy data). If many of these data are missing the outcome of a study may be inconclusive.	>=10%	Check the assumption, check for different Therapeutic Areas, understand distribution
% or number of subjects with premature trial drug discontinuation	Monitor and mitigate the risk associated with a significant proportion of Subjects experiencing interruptions in their treatment regimen what can result in not enough Subjects or poor statistical significance	>=5%	Check the assumption, check for different Therapeutic Areas, understand distribution
% or number of lost to follow-up subject	Dropouts due to loss to follow-up are a subset of all dropouts from the trial. The proportion of these dropouts is critical in outcome studies as regulators are concerned with a high dropout rate that the dropouts may be disproportionate due to poor outcomes caused by the investigational product.	3-4% MAX	Check the assumption, check for different Therapeutic Areas, understand distribution

% or number of incorrectly randomized or stratified subjects	Ensure the accuracy and reliability of Subject randomization/stratification in clinical trials or research studies, thereby enhancing the validity of study findings and supporting evidence-based decision-making.	>=3%	Check the assumption, check for different Therapeutic Areas, check distribution and probability to meat the requirement
% or number of subjects with AEs/SAEs of special interest	Monitoring safety is critical in a clinical trial. By monitoring the number of (S)AEs reported per participant visit and comparing between sites, it is possible to detect sites that are either reporting unusually high or low levels of (S)AEs. These outlier sites need investigation as it may be that the protocol is not being followed (e.g., the same AE is being reported multiple times, multiple AEs are being reported as one AE), or that the trial participants at those sites are generally healthier (or less healthy) than at other sites. Investigation is needed to determine if further follow-up and correction is required.	>=50%	Check the assumption, check for different Therapeutic Areas, understand distribution
Rate of AE per patient		Very Specific, Requires additional Analysis for TA	Check the assumption, check for different Therapeutic Areas, understand distribution
Rate of SAE per patient		Very Specific, Requires additional Analysis for TA	Check the assumption, check for different Therapeutic Areas, understand distribution
Long Mean Time for Response to Queries *(in case of access to Audit Trail)*	A KRI checking whether the sites are always up to date with respect to answering queries generated by the EDC edit checks, prevent delayed statistical analysis.	7	Check the assumption, check for different Therapeutic Areas, Different type of Queries (Automatic/Manual), understand distribution
Long mean time for data entry	Delays in data entry into the data capture system may jeopardize the Subject's safety and / or have a negative impact on the time of database closure / availability of the study results	5	Check the assumption, check for different Therapeutic Areas, understand distribution

2. Data Behavior and Statistical Distribution: We will analyze the patterns in clinical trial data to understand how it behaves statistically. This will help us develop better quality control methods that are specific to the unique aspects of clinical trials.

3. Duplication Probability in Patient Data: We will investigate the likelihood of having duplicate patient records in the data collected during clinical trials. Using advanced statistical techniques, we aim to identify and address issues related to data integrity, participant safety,

and data management processes. Additionally, we will explore and evaluate different statistical algorithms to efficiently detect duplicate patient data, further improving data quality assurance measures.

Our overall goal is to revolutionize how clinical trial data is managed, providing valuable insights and solutions to researchers, clinicians, health authorities, and other stakeholders. By doing so, we hope to enhance the quality and reliability of outcomes in clinical trials, contributing to advancements in medical research and public health.

Research Proposal

Cyntegrity has initiated Project 1 with the primary objective of developing statistical methods for the review and control of quality tolerance limits in clinical trials. The application of these statistical methods necessitates the analysis of extensive datasets to comprehend general trends, behavioural patterns, and data distribution, among other aspects.

We will focus on analyzing information related to subjects, including visits, adverse events, lost to follow-ups and outcomes/endpoints reporting processes. This analysis will encompass various factors such as the number of escalations, their dynamics, top limits (maximum and minimum number of items), and the average duration of a case (period when number of cases increased desired level or was higher than average study level or some pre-defined benchmark).

Our investigation aims to identify common trends in similar clinical trials across therapeutic areas, phases, administration methods of the investigational product, time considerations (for a site) in the study, and potentially other relevant factors. The overarching goal of the project is to ascertain the most prevalent statistical distribution for a specific quality tolerance limit and establish benchmarks for key parameters.

It is crucial to note that the project explicitly excludes the assessment of the effectiveness of the investigational product; the focus remains on statistical aspects related to quality tolerance limits associated with study management processes.
Project 2: Cyntegrity is currently engaged in a project with the primary objective of identifying a minimal set of parameters that can facilitate the detection of "Duplicated Subject".

Scientific Abstract

Brief Project Background and Statement of Project Significance

Background:

Clinical trials represent a cornerstone in advancing medical knowledge and healthcare interventions. As the volume and complexity of clinical trial data continue to grow, there is a critical need to understand and optimize the quality, behavior, and integrity of this data. The proposed research project by Cyntegrity aims to comprehensively investigate quality tolerance limits, data behavior, and duplication probability in clinical trials across diverse therapeutic areas, phases, and trial types. The background of this project stems from the increasing challenges associated with managing and interpreting vast datasets within the context of clinical trial quality. By leveraging advanced data science techniques and statistical methodologies, the research seeks to unravel patterns, benchmarks, and probabilities that significantly impact the reliability and precision of clinical trial outcomes. The significance of this project is multiform. It addresses critical needs in current practices and contributes to the advancement of clinical trial research and data management in the following ways.

Objective:

Firstly, the clinical trial design will be optimized. Understanding benchmark quality tolerance limits will enable the optimization of clinical trial designs, leading to more efficient studies. Secondly, it enhanced the data quality and reliability. Insights into data behavior and statistical distributions will enhance the accuracy and reliability of predictive models, fostering more dependable clinical trial data. Thirdly, the investigation into duplication probability directly impacts patient data integrity, minimizing risks associated with duplicate records and improving overall data management practices. Overall, by disseminating findings through publications and collaboration with the scientific community, the project aims to contribute to the advancement of industry standards, fostering continuous improvement and innovation in quality control of clinical trial research.

Study Design:

The study is composed of data management and statistical analysis. Study data will be categorized according to phases and therapeutical areas. Based on the requirements of quality tolerance limit such as rate of adverse event, lost to follow-up, missing endpoint, etc., data from various studies is converted into a standardized format which is used in calculating the matrices on site, country, and study level. To understand the best model for study in different stages and therapeutical areas, multiple distributions are tested on matrices and summarize the best parameter set for each group of study type.

Participant:

In the context of the two projects, we are going to investigate the clinical trial data quality and integrity. Therefore, the project is not limited to a specific patient group. To develop a solid solution, the project would like to include patients in different ages, ethnicities, races, and medical history backgrounds, so that the developed algorithm is applicable to different types of studies.

Primary and Secondary Outcome Measure(s):

The primary outcome of the project is to determine the best distribution that either the site or center monitor can follow during clinical trials. Comparing the selected distribution with the study data can assess whether a study is under control or not. Regarding patient duplication detection the primary outcome is to assess the similarity between the given patient and all the patients in the database. For both topics, we will not take the secondary endpoint into account because the purpose of the research is to develop a new method of data management. We are not going to investigate/compare etc. product efficacy/safety etc.

Statistical Analysis:

The goodness of fit and patient similarity is evaluated by numeric analysis approach, Anderson-Darling statistic, and significant test. By patient similarity, the statistic test defines the threshold of identity. Regarding the quality tolerance limit, some matrices depend on the number of study days. The numeric analysis approach can identify the pattern of the time series variable and then use statistical tests again to evaluate the goodness of fit. Moreover, considering the hypothesis of the best-fit distribution, the researchers assess if the quality tolerance limits are under the same scenario as the best-fit distribution.

Use of Information for Generalizable Scientific and/or Medical Knowledge:

The insights gained from this research will be disseminated through peer-reviewed publications, conference presentations, and collaborative engagement with the scientific community. By sharing methodologies, findings, and recommendations, the information obtained will materially enhance knowledge and experience in clinical trial quality control. This knowledge, in turn, will inform future research endeavours, influence data management practices, and contribute to the collective understanding of how to optimize clinical trials for the betterment of science and public health.

References:

ICH E6 (R3) Guideline on good clinical practice (GCP)_Step 2b (europa.eu)
ICH: E6 (R2): Guideline for good clinical practice - Step 5 (europa.eu)
ICH guideline E8(R1) Step 2b on general considerations for clinical studies (europa.eu)

Specific Aims of the Project

Topic 1 (QTL):

Examine prevalent or unique subject related data and benchmarking for subject-related quality tolerance limits employed in clinical trials. Gain insights into the types of distributions, distribution parameters (with recommendations), and assess whether there are specific patterns related to therapeutic areas or study phases.

Topic 2 (Duplicated records):

Identify a set of subject characteristics that can facilitate the detection of Duplicated Subject.

What is your Study Design? Please select one of the following options

o Individual trial analysis

o [✓ Meta-analysis (analysis of multiple trials together)]

o Methodological research

o Other

What is the purpose of the analysis being proposed? Please select all that apply.

New research question to examine treatment effectiveness on secondary endpoints and/or

within subgroup populations

o New research question to examine treatment safety

o Confirm or validate previously conducted research on treatment effectiveness

o Confirm or validate previously conducted research on treatment safety

o Preliminary research to be used as part of a grant proposal

o [✓ Summary-level data meta-analysis]

□ Meta-analysis using only data from the YODA Project

  Meta-analysis using data from the YODA Project and other data sources

o Participant-level data meta-analysis

  Meta-analysis using only data from the YODA Project

  Meta-analysis using data from the YODA Project and other data sources

o Develop or refine statistical methods

o Research on clinical trial methods

o Research on comparison group

o Research on clinical prediction or risk prediction

o Other

New research question to examine treatment effectiveness on secondary endpoints and/or within subgroup populations

o New research question to examine treatment safety

o Confirm or validate previously conducted research on treatment effectiveness

o Confirm or validate previously conducted research on treatment safety

o Preliminary research to be used as part of a grant proposal

o Summary-level data meta-analysis

  Meta-analysis using only data from the YODA Project

  Meta-analysis using data from the YODA Project and other data sources

o Participant-level data meta-analysis

  Meta-analysis using only data from the YODA Project

  Meta-analysis using data from the YODA Project and other data sources

o Develop or refine statistical methods

o Research on clinical trial methods

o Research on comparison group

o Research on clinical prediction or risk prediction

o Other

  Please explain

Research Methods

Data management:
Cyntegrity will create a data base for combining data from YODA and Cyntegrity archive. The data from two data sources is planned to be standardized into a unified format which includes only the information required for the analysis. The customized database can be SQL database on azure

depend on the accessibility of YODA data.

Statistical Analysis plan:
Topic 1:
The quality tolerance limit monitors clinical trials data through numeric analysis. The numeric analysis is conducted on derived metrics such as percentage of dropout, rate of adverse event etc. Metrics will be converted into density-based distribution and will be observe on site level and exam through the 108 types of continuous distribution's probability density function to figure out the best match parameter set which is with less deviation of curve area between real-world data and simulation. Moreover, we quantify the goodness of fit by using Anderson-Darling statistic on continuous data. The selected study will be labelled by their therapeutical area and phases and classify. The hypothesis is that studies from different therapeutic area and phase can show different patterns on the quality tolerance limit.

Topic 2:
Duplicated patients detection is conducted by crossing studies and subgrouping patient features such as medical history, vital sign, geolocation and other physical examinations. The features from patient will be converted into multi-dimensional matrix. Similarity is evaluated by the dot product of matrices.

Result and conclusion:
Regarding to combining the analysis result, we will take the parameter set from both sides and use the weighting to come out a combined parameter set. And then we will validate it with visualization.

Software to be used

o [✓ Python]

o R

o RStudio

o STATA

o Open Office

o I am not analysing participant-level data / I plan to use another secure data sharing platform ♣ Please clarify

Data Source and Inclusion/Exclusion Criteria to used to define the patient sample for your study

The analysis will be based on three data sources which are Cyntegrity owned data, open-source clinical trials data and Data from YODA. The study data from phase 1 and 4 will be excluded. During data cleaning stage, the records with unclear date or missing information will not be considered.

Primary and Secondary Outcome Measure(s)

Topic 1.

The primary outcome of the project is to determine the best guidance that either the site or center monitor can follow during clinical trials. The guidance is distributions which can explain the quality tolerance limit and statistically significant represent the distribution of metrics calculated by a study from each subgroup. For instance, the center monitor can know from the guidance that their study is overreporting or underreporting adverse events. Furthermore, according to the guidance, the center monitor can use it as a model to adjust the parameters that could better fit their study requirement for example, a higher adverse event report rate at an early stage is expected.

Topic 2.

The primary outcome of the patient duplication detection is to assess the similarity between the given patient and all the patients in the database. Therefore, the optimal outcome is a score of similarity taking the vital sign, geolocation, and medical history into consideration. In addition, another probability score can indicate the rarity of getting this similarity score.

For both projects, we will not take the secondary endpoint into account because the purpose of the research is to develop a new method of data management. We are not going to investigate/compare etc. product efficacy /safety etc.

Main Predictor/Independent Variable and how it will be categorized/defined for your study

Topic 1. The analysis will be based on three data sources which are Cyntegrity owned data, open-source clinical trials data and Data from YODA. Study data will all be anonymized and categorized by the therapeutic area and study phase.

Topic 2. Patient data will be anonymized. The predictor variables are vital sign, site location, medical history and concomitant medication, which will not be categorized but one-hot encoding and convert into muti-dimensional vector for the similarity calculation algorithms.

Other Variables of Interest that will be used in your analysis and how they will be categorized/defined for your study

NA

Statistical Analysis Plan

Topic 1:

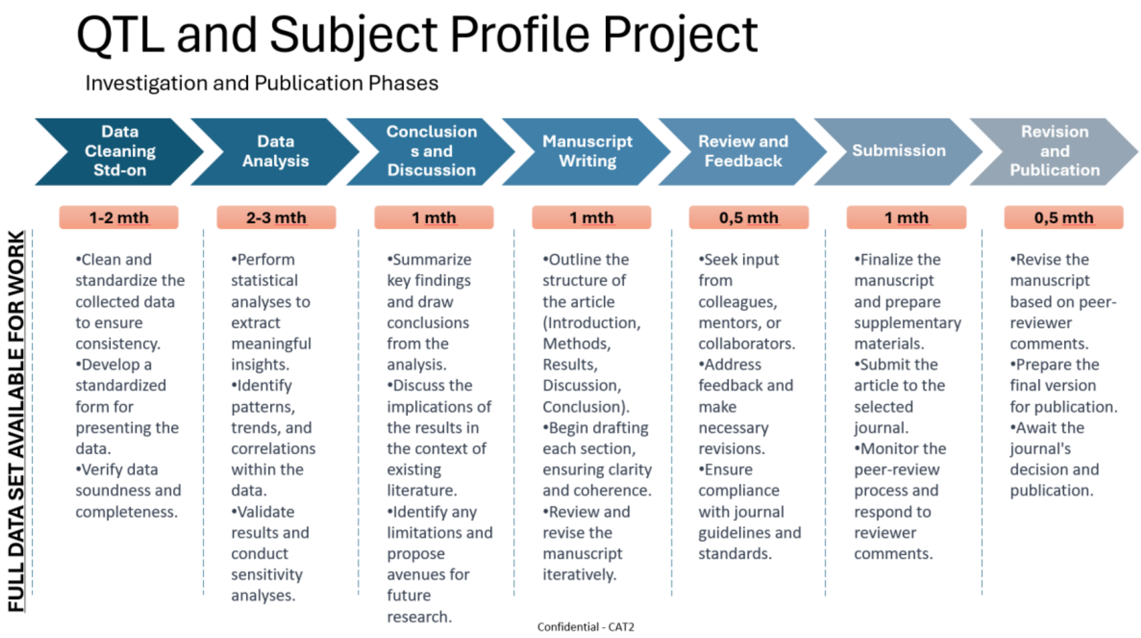
The quality tolerance limit monitors clinical trials data through numeric analysis. The numeric analysis is conducted on derived metrics such as percentage of dropout, rate of adverse event etc. Metrics will be converted into density-based distribution and will be observe on site level and exam through the 108 types of continuous distribution's probability density function to figure out the best match parameter set which is with less deviation of curve area between real-world data and simulation. Moreover, we quantify the goodness of fit by using Anderson-Darling statistic on continuous data. The selected study will be labelled by their therapeutical area and phases and classify. The hypothesis is that studies from different therapeutic area and phase can show different patterns on the quality tolerance limit.

Topic 2:

Duplicated patients' detection is conducted within and across studies and subgrouping patient features such as medical history, vital sign, geolocation and other physical examinations. The features from patient will be converted into multi-dimensional matrix. Similarity is evaluated by the dot product of matrices.

Project Timeline

1. Data Cleaning Std-on (1-2 month):
   - Clean and standardize the collected data to ensure consistency.
   - Develop a standardized form for presenting the data.
   - Verify data soundness and completeness.
2. Data Analysis (2-3 months):
   - Perform statistical analyses to extract meaningful insights.
   - Identify patterns, trends, and correlations within the data.
   - Validate results and conduct sensitivity analyses.
3. Conclusions and Discussion (1 month):
   - Summarize key findings and draw conclusions from the analysis.
   - Discuss the implications of the results in the context of existing literature.
   - Identify any limitations and propose avenues for future research.
4. Manuscript Writing (1 month):
   - Outline the structure of the article (Introduction, Methods, Results, Discussion, Conclusion).
   - Begin drafting each section, ensuring clarity and coherence.
   - Review and revise the manuscript iteratively.
5. Review and Feedback (0.5 month):
   - Seek input from colleagues, mentors, or collaborators.
   - Address feedback and make necessary revisions.
   - Ensure compliance with journal guidelines and standards.
6. Submission (1 month):
   - Finalize the manuscript and prepare supplementary materials.
   - Submit the article to the selected journal.
   - Monitor the peer-review process and respond to reviewer comments.
7. Revision and Publication (0.5 month):
   - Revise the manuscript based on peer-reviewer comments.
   - Prepare the final version for publication.
   - Await the journal's decision and publication.



QTL and Subject Profile Project

Investigation and Publication Phases

FULL DATA SET AVAILABLE FOR WORK

Data Cleaning Std-on	Data Analysis	Conclusions and Discussion	Manuscript Writing	Review and Feedback	Submission	Revision and Publication
1-2 mth	2-3 mth	1 mth	1 mth	0,5 mth	1 mth	0,5 mth
•Clean and standardize the collected data to ensure consistency. •Develop a standardized form for presenting the data. •Verify data soundness and completeness.	•Perform statistical analyses to extract meaningful insights. •Identify patterns, trends, and correlations within the data. •Validate results and conduct sensitivity analyses.	•Summarize key findings and draw conclusions from the analysis. •Discuss the implications of the results in the context of existing literature. •Identify any limitations and propose avenues for future research.	•Outline the structure of the article (Introduction, Methods, Results, Discussion, Conclusion). •Begin drafting each section, ensuring clarity and coherence. •Review and revise the manuscript iteratively.	•Seek input from colleagues, mentors, or collaborators. •Address feedback and make necessary revisions. •Ensure compliance with journal guidelines and standards.	•Finalize the manuscript and prepare supplementary materials. •Submit the article to the selected journal. •Monitor the peer-review process and respond to reviewer comments.	•Revise the manuscript based on peer-reviewer comments. •Prepare the final version for publication. •Await the journal's decision and publication.

Confidential - CAT2

Dissemination Plan

Publications/Presentations: We will publish results in Applied Clinical Trials, present results in different conferences such as DIA, Phuse, Scope and SCDM

Auditory: Study Risk Management, Centralized Monitoring Management
The thresholds (medium and high) and distributions of Quality Tolerance limits are of paramount significance in the realm of clinical study data quality control. These metrics (provided they are derived from meticulously evaluated and mathematically validated data related to similar trials) are indispensable for the formulation of a Sponsor's quality strategy and for the management of ongoing clinical trials, . The development of a quality strategy typically falls within the purview of Study Risk Management, with the ongoing oversight of studies from a risk management perspective being the responsibility of the Centralized Monitoring management team. In addition to mitigating and managing study risks, Centralized Monitoring teams are typically entrusted with overseeing study data fraud detection, which means that patient data duplication serving as a pivotal and challenging checkpoint in this regard. The findings of Cyntegrity's research endeavors are slated for dissemination and presentation through relevant publications or platforms (Society for Clinical Trials (sctweb.org), Journal of Clinical Monitoring and Computing (springer.com)) , and conferences (DIA, Phuse, Scope and SCDM) aimed at the aforementioned target audiences.

Bibliography

ICH E6 (R3) Guideline on good clinical practice (GCP) Step 2b (europa.eu)
ICH: E 6 (R2): Guideline for good clinical practice - Step 5 (europa.eu)
ICH guideline E8(R1) Step 2b on general considerations for clinical studies (europa.eu)

Supplementary Material