GENINVO™

Data Science & Analytics

NLP TEXT SIMILARITY

Quality Check for DocQc Table-41

www.geninvo.com

North America | UK | Europe | India

Business Need

- The clinical study report (CSR) is a crucial document in the drug development and regulatory submission process that describes the background, methods, analysis, and results of a clinical trial. In the regulatory approval process and premarket regulatory review, the quality of the CSR is important. The quality check (QC) process verifies data reported in the CSR against source data in order to ensure that the results presented are accurate, statements are supported by citations, and conclusions align with the results. CSR-QC is therefore integral and the most crucial step of the CSR preparation process to ensure accuracy and quality. Improper CSR preparation can not only result in additional questions from regulatory authorities, but it can also negatively impact trial credibility.

Challenges

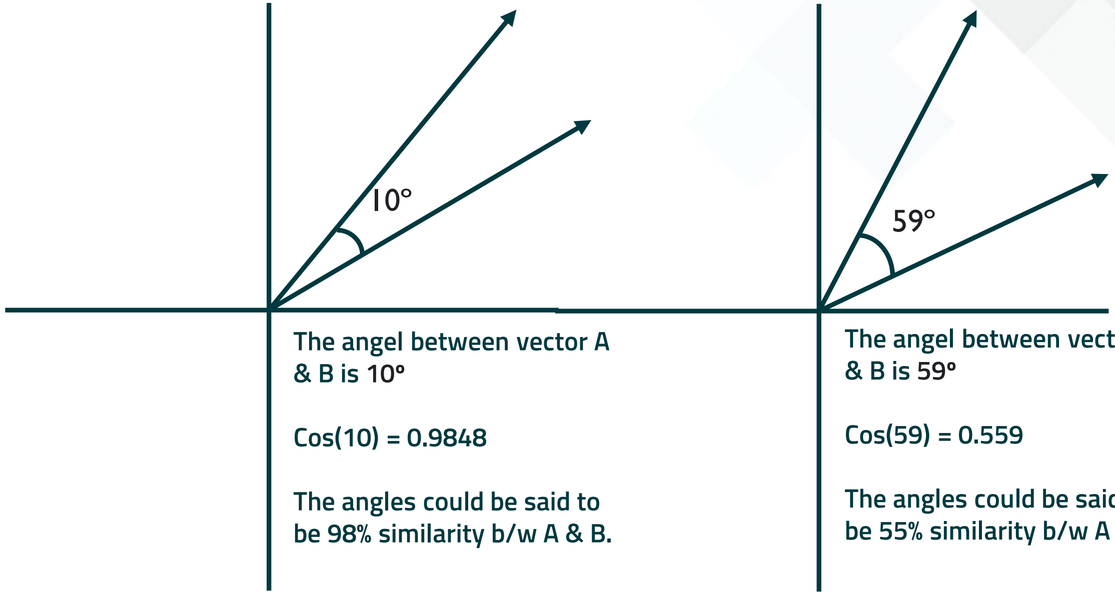
CSR-QC is one of the major contributors to delayed reporting of trial data. The current manual QC process is inadequate and poses significant challenge regarding costs, time, quality and resource management.

- An extensive manual quality check was involved in the complex study report involving extensive numerical data
- There is a high likelihood of observational errors during the CSR review cycle, which delays the overall turnaround time for producing documents

The most critical and time-consuming checkpoint for a CSR-QC is determining whether the information reported in the In-text are accurate based on the Source Tables.

Solution Methodology

- A thorough analysis of CSR was conducted to understand how data could be extracted based on table reference from in-text as well as source tables
- NLP techniques used for data processing and cleaning
- Cosine similarity is used to compare in-text data with source table data. Cosine Similarity is a measurement that quantifies the similarity between two or more vectors. The cosine similarity is the cosine of the angle between vectors. The vectors are typically non-zero and are within an inner product space. Cosine Similarity is a value that is bound by a constrained range of 0 and 1. As the cosine similarity measurement gets closer to 1, then the angle between the two vectors A and B is smaller. The images below depict this more clearly.



The angel between vector A & B is 10°

Cos(10) = 0.9848

The angles could be said to be 98% similarity b/w A & B.

The angel between vector A & B is 59°

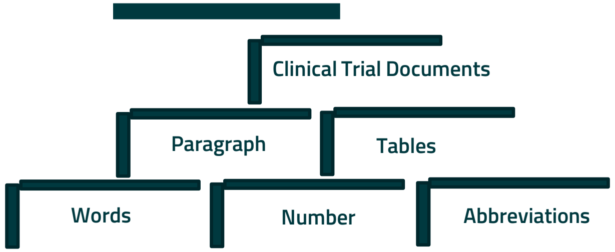
Cos(59) = 0.559

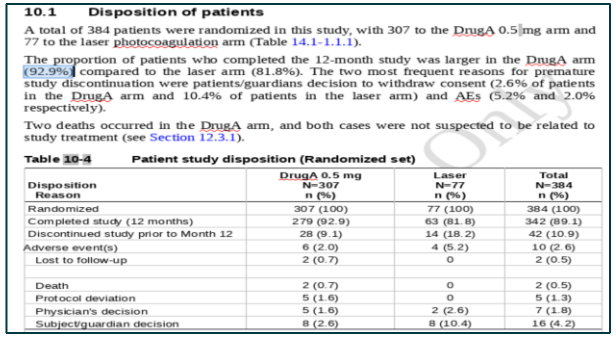
The angles could be said to be 55% similarity b/w A & B.

Based on pre-defined threshold, cosine similarity score is divided into three category to know more about the matches

- Exact Match – cosine score value range b/w 0.99 to 1.
- Partial Match – cosine score value range b/w 0.8 to 0.98.
- No Match - cosine score value is 0.

NLP Cosine Similarity Flow Map





10.1 Disposition of patients

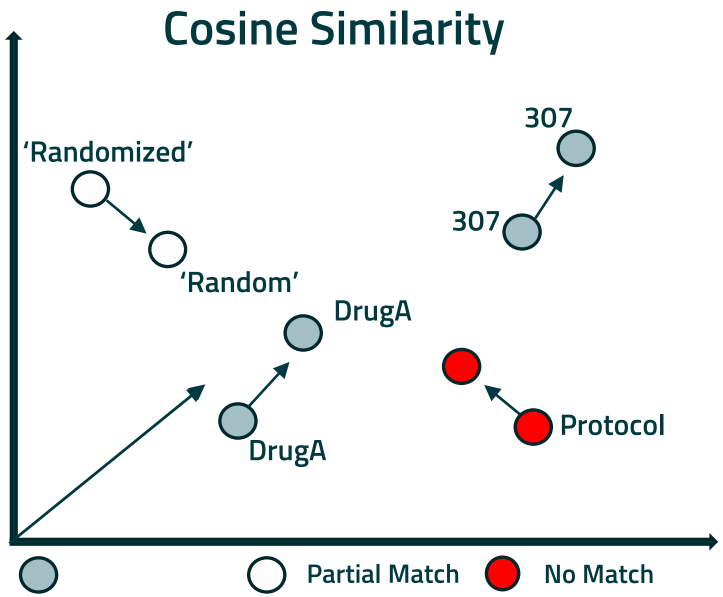
A total of 384 patients were randomized in this study, with 307 to the DrugA 0.5 mg arm and 77 to the laser photocoagulation arm (Table 14.1-1.1.1).

The proportion of patients who completed the 12-month study was larger in the DrugA arm (92.9%) compared to the laser arm (81.8%). The two most frequent reasons for premature study discontinuation were patients/guardians decision to withdraw consent (2.6% of patients in the DrugA arm and 10.4% of patients in the laser arm) and AEs (5.2% and 2.0% respectively).

Two deaths occurred in the DrugA arm, and both cases were not suspected to be related to study treatment (see Section 12.3.1).

Table 10-4 Patient study disposition (Randomized set)

Disposition Reason	DrugA 0.5 mg N=307 n (%)	Laser N=77 n (%)	Total N=384 n (%)
Randomized	307 (100)	77 (100)	384 (100)
Completed study (12 months)	279 (92.9)	63 (81.8)	342 (89.1)
Discontinued study prior to Month 12	28 (9.1)	14 (18.2)	42 (10.9)
Adverse event(s)	6 (2.0)	4 (5.2)	10 (2.6)
Lost to follow-up	2 (0.7)	0	2 (0.5)
Death	2 (0.7)	0	2 (0.5)
Protocol deviation	5 (1.6)	0	5 (1.3)
Physician's decision	5 (1.6)	2 (2.6)	7 (1.8)
Subject/guardian decision	8 (2.6)	8 (10.4)	16 (4.2)


Cosine Similarity

Source Data

Data Processing

Libraries
gensim
nltk
torchtext
spacy

Pre-Processing

1. Read docx/pdf file
2. Convert to json format
3. Extract paragraph & Table

Processing

1. Convert to lower case
2. Remove Stop-words
3. Remove Symbols
4. Stacking
5. list

Algorithm

Results

Exact Match : 307 with original number: 307

Partial Match : Randomized with original Random.

No Match : No match for word Protocol

Business Impact

- Up to x% reduction in average quality check time
- Significant cost savings and increased throughput in operations
- 100% SLA adherence leading to no penalties
- Achieving the much-needed first mover's advantage
- Minimal loss to pharma companies

GENINVO™

Ideas
Delivered
Intuitively

Thank You!

www.geninvo.com | Reach out to us info@geninvo.com