FACTORS PREDICTING QUERY OUTCOMES IN OBSERVATIONAL STUDIES SETTING

Fabio Ferri[1], Alessandra Mignani[1], Simone Schena[1], Giulio Mazzarelli[1], Lucia Simoni[1], Alessandra Ori[1]

[1]IQVIA Solutions Italy s.r.l.

BACKGROUND

- Managing queries (defined as requests for clarification or for resolving data discrepancies on an electronic data capture system) in observational studies remains challenging, as during data cleaning numerous queries may be generated.
- Frequent query rejections due to unclear or unacceptable justifications for data discrepancies significantly delay study completion and increase workload for research staff, but also for the Data Manager (who is responsible for reviewing the consistency and completeness of data during the cleaning process).
- **The aim of this project was to investigate the extent to which the approval or rejection of a query in observational studies can be predicted by analyzing its textual content and contextual information.**

METHODS

- Data from observational studies (including both multi-country and local studies) were analyzed using SAS Enterprise Guide. Studies were selected to embody diverse therapeutic areas and designs to capture variability in query patterns.
- The analysis involved an integrated process of data harmonization across datasets, classification of queries using predefined descriptors (i.e., queries raised on forms completed at scheduled visits or on forms completed at any time during the study; queries raised manually by the data manager after performing data checks of greater complexity vs simpler queries raised automatically by the system), and feature extraction from query text through linguistic analysis. Extracted features included text complexity metrics (query text length, word count, average word length), action word indicators (variables for the presence/absence of terms like "check", "verify"), and content indicators (references to ranges, dates).
- A multicollinearity assessment was conducted using correlation diagnostics to identify and address potential collinearity issues among predictor variables.
- As described in **Figure 1**, a **machine learning**[1] approach using stratified data splitting (70% training, 30% testing) was employed to develop a logistic regression model. The variable selection was conducted on the training set using a stepwise approach based on significance level ($p < 0.05$). This approach helped to address bias in parameter estimates on the test set, reduce the risk of overfitting, enhance the model's ability to generalize to unseen data, and provide a more accurate assessment of the model's performance.

RESULTS

- Out of 28175 queries, 7.5% (n=2119) were rejected. Among queries raised manually by the data manager (n=10376, 36.8%), 1362 (13.1%) were rejected.
- As reported in **Table 1**, out of 26056 approved queries 39.70% requested to update certain information, whereas only 19.35% of the rejected queries (N=2119) included a request to update certain information.
- An increase in the average word length of the query text was associated with higher odds of rejection (OR=1.34, 95% CI: 1.15 – 1.57), as shown in Table 2.
- **Table 2** shows that the request to provide clarification had a rejection odds ratio (OR) of 8.73 (95% CI: 5.17-14.76). Queries containing a request referring to time and dates inconsistencies (OR=1.76, 95% CI: 1.44-2.15) or to inspect consistency of values and ranges (OR=1.64, 95% CI: 1.20-2.24) increased the odds of rejection.
- **Table 2** shows that queries requesting to "update" data had a decrease in odds of rejection (OR=0.76, 95% CI: 0.58-0.99), as well as the requests directly addressed to the person in charge in queries resolution using personal pronouns (i.e., "You stated that…") in the query text (OR=0.64, 95% CI: 0.53-0.78).
- The model identifies 54% of actual rejected queries and 72% of actual approved queries.

CONCLUSIONS

- **The feature-based machine learning approach, utilizing logistic regression with feature selection, indicates that specific text elements can help predict query rejection in observational studies.**
- **The request to provide clarifications, checking laboratory ranges and dates, and complex language with overly long words can all be challenging for investigators.**
- **The request to update data does not have a relevant impact on the query rejection.**
- **This approach combines predictive power with interpretability, offering practical guidance for optimizing data management workflows and prioritizing review efforts in the real-world setting.**

Figure 1: Work-flow for the the project activities

Data Import and Harmonization

Importing datasets into SAS and **harmonizing** them to ensure consistency in format and structure across all data sources.

Data Manipulation

Filtering the dataset and **engineering features** from the query text to enhance the predictive power of the model.

Exploratory Data Analysis

Understanding the **distribution, relationships**, and **patterns within the data**, and identifying key features from the query text.

Variables selection

A logistics regression witih **stepwise selection** was applied on the training set.

Estimate of parameters

Using selected variables, **logistic regression** on the test set was implemented and Odds Ratio (OR) for queries rejection were estimated.

Model evaluation

The evaluation of the model's performance was conducted on the test set, focusing on the trade-off between **sensitivity and specificity.**

MACHINE LEARNING APPROACH

Table 1: Descriptive statistics of main significant action word indicators by Query Outcome

Query text contains the request…	Approved Queries* (N=26056)	Rejected Queries** (N=2119)
to **check** information	11689 (44.86%)	1371 (64.70%)
to **update** information	10355 (39.74%)	410 (19.35%)
to **provide** information	125 (0.48%)	104 (4.91%)
to inspect consistency of **values or ranges**	1748 (6.71%)	184 (8.68%)
referring to **Date/Time** inconsistencies	7445 (28.57%)	936 (44.17%)
directly addressed to the person in charge	9972 (38.27%)	532 (25.11%)

*Percentages were computed by overall queries approved.
**Percentages were computed by overall queries rejected.

Table 2: Odds Ratios and 95% Confidence Intervals for queries rejection (Test Set)

Query features	Odds Ratio	95% Confidence Interval
Request to **provide** information (Yes vs No)	8.73	(5.17 - 14.76)
Request referring to **Date/Time** inconsistencies (Yes vs No)	1.76	(1.44 - 2.15)
Request to inspect consistency of **values** or **ranges** (Yes vs No)	1.64	(1.20 - 2.24)
Request to **check** information (Yes vs No)	1.38	(1.06 - 1.79)
Forms compiled at **scheduled visits** vs forms completed at **any time point**	1.38	(1.14 - 1.67)
Average Word Length	1.34	(1.15 – 1.57)
Queries raised **automatically vs manually**	0.37	(0.30 - 0.46)
Request **directly addressed** to the person in charge (Yes vs No)	0.64	(0.53 - 0.79)
Request to **update** information (Yes vs No)	0.76	(0.58 - 0.99)

Significant and relevant variables for the analysis have been reported in the table.

For any additional information on this content please contact Alessandra Ori at alessandra.ori@iqvia.com or scan this code

1. Chai KEK, Anthony S, Coiera E, Magrabi F. Using statistical text classification to identify health information technology incidents. Journal of the American Medical Informatics Association. 2020;27(1):62-70.

64 SIMPOSIO AFI RIMINI 11·12·13 GIUGNO 2025
Associazione Farmaceutici Industria Società Scientifica