Paper IC04

Novel Analytics for Risk-Based Monitoring: Harnessing Natural Language Processing for Query Data to Enhance and Optimize Data Quality

Lawrence Dennison-Hall, Roche Products Ltd, Welwyn Garden City, United Kingdom
Mariia Lapaeva, F. Hoffmann-La Roche, Basel, Switzerland

ABSTRACT

In healthcare analytics, unstructured free-text often contains vital information, posing challenges for traditional processing methods. Our initiative pioneers a novel solution to proactively detect data quality concerns within previously unanalyzed textual query data.

We leverage two machine learning (ML) frameworks: XGBoost for analyzing query text and responses for inconsistencies and pre-trained SBERT for detecting reopened queries. Our aim is to preemptively address data quality concerns and optimize resource allocation based on risk-based monitoring (RBM) principles.

The models achieve accuracy of 89.6% and 90% respectively (test sets: ~10000 queries). Integrated into Roche's Data Quality Oversight exploratory tool for over 450 studies, these models provide real-time insights and offer risk-based solutions for identifying data quality challenges. Our qualitative user feedback underscores the transformative potential of ML. This ultimately enables us to offer deeper insights into potential growth in the data quality backlog, facilitating the proactive mitigation of data quality issues.

INTRODUCTION

In pharmaceutical research and development, the evolution of risk-based monitoring (RBM) has ushered in a paradigm shift, redefining the approach to clinical trial oversight [1]. Traditionally, clinical trial monitoring has involved extensive on-site visits and data reviews, often resulting in resource-intensive processes. However, the incorporation of RBM principles have revolutionized this landscape by focusing on a targeted, risk-based strategy to optimize resource allocation and enhance data quality. RBM principles such as Quality-by-Design (QbD) and Critical-To-Quality (CTQ) are also creating a demand for insights related to prospective risk detection to aid conscious decision making around fit-for-purpose control strategies. One focus area for risk detection is the Electronic Data Capture (EDC) system used in clinical trials and the trends emerging from data that are being queried for quality control.

One of the key advances that has the potential to improve RBM strategies, specifically for EDC queries, is the automation of processes and the integration of machine learning (ML) and natural language processing (NLP) methods. The growth of unstructured data, particularly in the form of free-text queries within healthcare systems, poses significant challenges for traditional data processing methods [2]. NLP methods, including word embeddings, offer a solution by enabling the extraction of meaningful insights from unstructured textual data. NLP, a field initially formalized in the 1950s, gained traction through key contributions from researchers such as Alan Turing and John McCarthy. The modern wave of NLP breakthroughs can be attributed to the emergence of word embeddings and neural network-based language models.

The introduction of word embeddings, notably Word2Vec by Mikolov et al. [3] and GloVe (Global Vectors for Word Representation) by Pennington et al. [4], marked a critical milestone in NLP. These models facilitated the representation of words as dense, continuous vectors, enabling algorithms to capture semantic relationships and meaning from textual data more effectively.

The revolutionary Bidirectional Encoder Representations from Transformers (BERT), proposed by Devlin et al. [5], sparked a new era in NLP. BERT introduced a pre-training strategy with bidirectional transformer models that allowed the model to understand the context of words within a sentence bidirectionally.

Building upon the success of BERT, Sentence-BERT (SBERT) was introduced by Reimers and Gurevych et al. in 2019 [6]. SBERT focuses specifically on converting sentences into fixed-dimensional embeddings that enable better semantic understanding and comparison between sentences. The innovation of SBERT lies in its ability to generate

sentence embeddings that capture nuanced semantic information from short pieces of text, making it useful for various NLP tasks. SBERT's ability to compare and understand sentences can be leveraged to analyze patient queries, adverse event reports, protocol deviations and medical records; enhancing the accuracy of information retrieval and decision-making in clinical settings. Overall, these technologies play a crucial role in expanding data-driven insights, accelerating research and ensuring informed decision-making in the pharmaceutical field.

In parallel, successfully advancing analysis in RBM, where data is represented in diverse formats including free-text, structured tables, numbers and images, necessitate the utilization of advanced algorithms and machine learning models. ML techniques, in particular decision trees and gradient boosting, can analyze complex non-linear relationships in patient data. Decision trees, developed by Ross Quinlan [7], can assist in understanding complex medical data which can include disease classification, drug response prediction, and patient risk assessment. XGBoost (an implementation of gradient boosting), introduced by Tianqi Chen and Carlos Guestrin in 2016 [8], has become a transformative tool due to its ability to process complex relationships in data. Its capabilities in classification and regression enable pharmaceutical companies to analyze various data formats - from free text to structured tables and numerical data. Specifically, in RBM, where data comes in various formats, XGBoost could play a crucial role in classifying potential data quality issues within textual queries. This capability is instrumental in ensuring data integrity and reliability in the pharmaceutical industry.

Therefore, the **goal of our paper** is to provide a generalized RBM solution, which is able to proactively detect data quality concerns within previously unanalyzed textual EDC query data, enabling insights that can help inform risk-based decision making across various clinical trial development activities. In our work, we are investigating it via three research questions:

RQ 1: Can Machine Learning models effectively identify reopened queries not directly found within electronic Case Report Forms (eCRF) metadata? Considering the time and resources involved in addressing each query, frequent reopening can significantly burden sites and signal inefficiency or misunderstanding. Analyzing trends and frequencies of reopened queries allows for early risk detection as well as actions aligned with RBM principles to reduce this load. Thus, the experiment's scope involves comparing text similarities between two queries to predict if one is a reopened query.

RQ 2: Can Machine Learning models determine if an eCRF data entry should be updated based on query and response text to help to reveal conflicting site activities? The formulation of query text and responses greatly influences quick and accurate issue resolution which prevents misunderstandings between parties and proactively addresses potential data quality concerns for critical eCRF fields. Therefore, the experiment aims to determine if a field requires an update by analyzing query and response texts followed by cross-referencing them with the eCRF data audit trail to confirm changes or identify contradictory behavior.

RQ 3: Can the proposed models be integrated into a comprehensive and intricate RBM solution capable of supporting cohesive RBM processes across diverse clinical trial development activities? Globally harmonized guidelines for Good Clinical Practice (GCP) mandate integrating quality and risk management into clinical trial design and operations to mitigate risks and detect or prevent issues early on. Not only developing ML models, but also embedding them into a unified system that proactively drives quality during study design and execution is crucial. Therefore, this question's scope involves outlining the high-level design ideas of an end-to-end RBM system to enhance monitoring efficiency by considering an open-source approach and aligning efforts across the industry.

Our paper is structured as follows: chapters 1 and 2 are dedicated to the description of our Machine Learning models: SBERT, designed for detecting reopened queries (RQ 1), and XGBoost, utilized for analyzing query text and response inconsistencies (RQ 2). These chapters are subdivided into sections covering materials and methods, results and discussion which provides a detailed overview of our methodologies and findings. In Chapter 3, we present our solution for integrating the proposed ML models into the Data Quality Oversight (DQO) tool. Additionally, we discuss a proposal for an end-to-end RBM system aimed at enhancing RBM efficiency, emphasizing an open-source approach (RQ 3). Finally, the Conclusion chapter summarizes our work which provides an overview of the contributions of our research.

CHAPTER 1. DETECTION OF REOPENED QUERIES

MATERIALS

Our goal is to capture potential reopened queries that are not found within EDC system metadata. We aimed to programmatically detect text similarity between queries, identifying reopened queries beyond EDC metadata (RQ 1). To achieve this, we used a pretrained SBERT ML model detailed in the subsequent Methods subchapter. This is used to obtain the similarity in between two sentences (or queries). For testing and optimizing the model’s performance, we used 8805 queries which were randomly selected across various therapeutic areas in clinical trials. These queries were manually labeled to distinguish reopened queries from non-reopened queries, forming 3628 groups, each with identical subjects, visits, CRF forms and fields.

METHODS

We designed a process leveraging the SBERT model, pretrained on the large language corpus [6] for sentence categorization based on similarity scores. The procedural steps involved were as follows:

1. **Data Preprocessing and Grouping:** Query data was preprocessed and grouped by specific subject, visit, form and field IDs to organize related queries. This grouping allows us to organize and analyze queries that are related to each other based on these identifiers. Once the queries are grouped, we sort them in descending order based on the date they were raised. This sorting ensures that the most recent queries appear at the top of each group.
2. **Sentence Embeddings with SBERT:** A pre-trained Sentence-Bert (SBERT) model was utilized to generate embeddings for each query. SBERT helped us to derive meaningful sentence embeddings which capture the semantic information of the queries.
3. **Cosine Similarity Calculations:** Once we obtained the embeddings, we utilized cosine similarity to compare the embeddings and calculate a similarity score ranging from 0 to 1. Cosine similarity measures the cosine of the angle between two vectors, providing a measure of similarity between them (Equation 1).

$$similarity(A,\ B) = cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}}$$

Equation 1. Cosine similarity in between two n-dimensional vectors, A and B, where A_i and B_i are the ith components of vectors A and B, respectively. The cosine similarity, $cos(\theta)$, is represented using a dot product and magnitude.

   This approach enabled us to quantify the semantic similarity between queries and obtain valuable insights for our analysis.
4. **Reopened Status Prediction:** A similarity score matrix was constructed using SBERT embeddings and cosine similarity calculations to predict the reopened status of queries within each group. If the similarity score between a query and the most recent one exceeded a threshold, the query was tagged as reopened. By applying this process, we managed to identify queries that are similar to the most recent one in each group and predict their reopened status.

The optimal threshold for determining reopened queries was identified based on the model's performance. This was achieved by assessing different thresholds followed by computing the accuracy score. The threshold of 76% was identified as optimal for maximizing the accuracy. Therefore, queries with a similarity score greater than 76% are considered to be reopened, while those with a score lower than the threshold are classified as not reopened. The overview of the process is represented on Figure 1.

The accuracy and balanced accuracy value (Equation 2) served as a benchmark for evaluating the overall effectiveness of our approach in terms of the correctness of reopened request detection.

$$Accuracy = \frac{TP+TN}{TP + FN+TN + FP}$$

$$Balanced\ Accuracy = \frac{1}{2}\left(\frac{TP}{TP + FN} + \frac{TN}{TN + FP}\right)$$

Equation 2. Accuracy and Balanced accuracy, where TP, TN, FP, FN refers to true positives, true negatives, false positives and false negatives respectively.

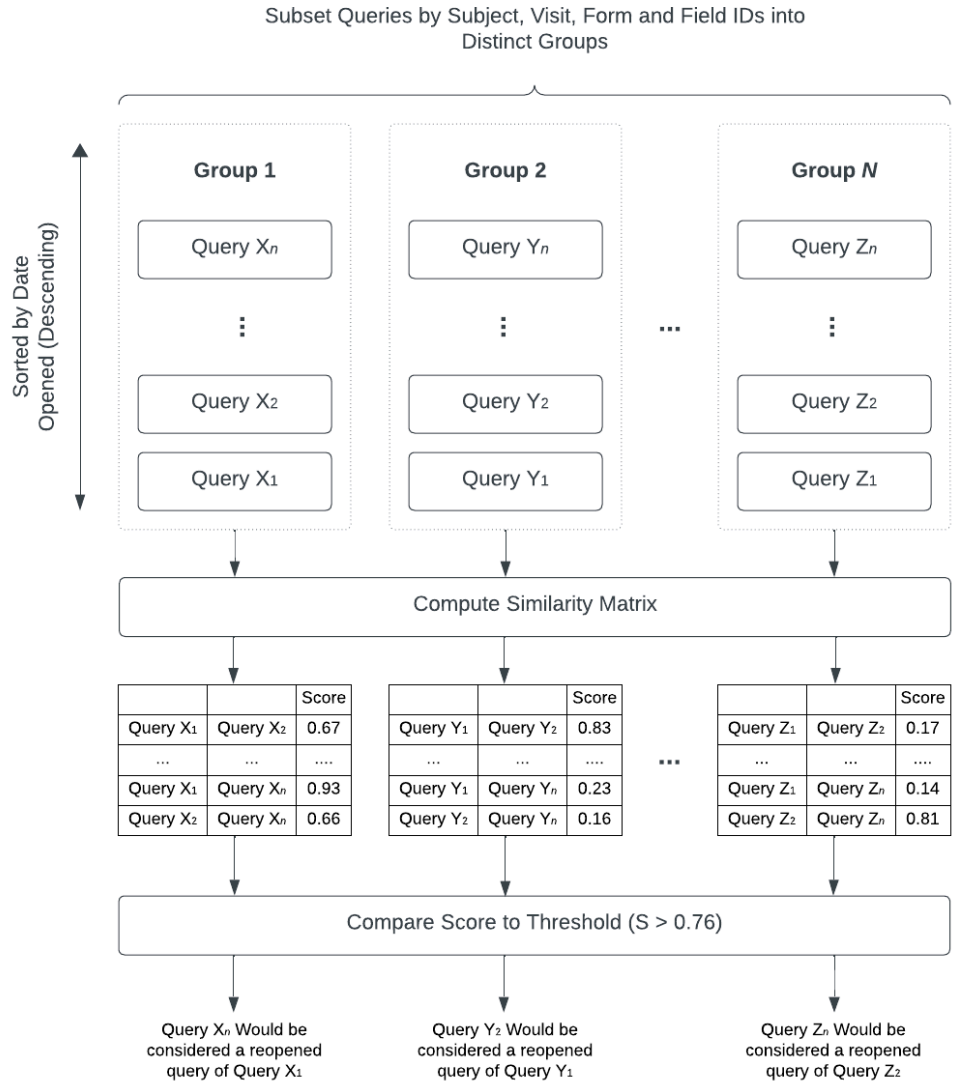

Figure 1. Overview of Determining Reopened Queries with Semantic Similarity

RESULTS

A similarity score threshold of 76% was identified as the point that yielded the most optimal results, achieving accuracy of 90% and balanced accuracy of 81.2% . The Confusion Matrix is represented on Figure 2.

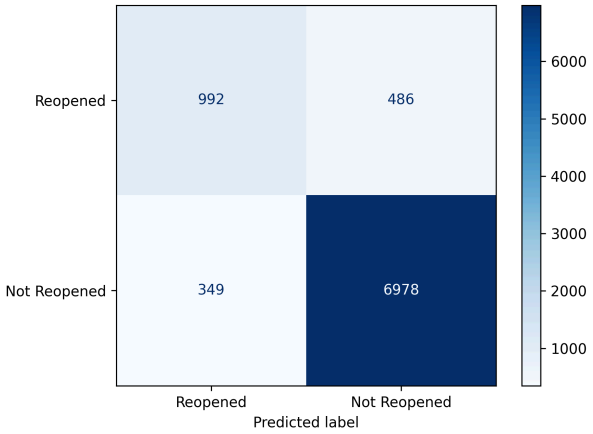

Figure 2. Confusion Matrix of SBERT for Detecting Reopened Queries. The predicted label of the model is shown along the X-axis which is compared to the ground truth on the Y-axis.

DISCUSSION

The results showcased the machine learning model's robustness in accurately categorizing queries, signifying its proficiency in identifying reopened queries not directly available within the eCRF, thereby effectively addressing RQ 1.

Importantly, however, we recognise the limitations of our approach. Firstly, although the SBERT model was chosen for its robustness, it was not pre-trained on EDC query data. This factor could lead to some inconsistencies in capturing dependencies within query words. Nevertheless, we assume that pre-training on a large language corpus could improve the generalisability and applicability of the model to different pharmaceutical companies while maintaining good prediction accuracy.

Secondly, the thresholds used in the analysis were set based on internal Roche EDC query data. This could potentially impact the performance of the model when applied to different query text styles. By setting this threshold, we aimed to strike a balance between correctly identifying reopened queries and minimizing misclassifications. However, it is important to note that the choice of threshold may vary depending on the specific requirements of the application and the trade-offs between different types of classification errors. Continuous evaluation and refinement of the threshold can help optimize the model's performance and ensure accurate categorization of queries as reopened or not reopened.

Additionally, we have observed a larger number of false positives relative to true positives in cases where the queries were actually reopened (Figure 2). This can be attributed to the presence of edge cases, which can pose challenges in predicting reopened queries with the same level of accuracy as non-reopened queries. Identifying and correctly categorizing these edge cases can be complex, as they may exhibit unique patterns or characteristics that deviate from the majority of queries. By further analyzing and understanding the specific characteristics of the false positives and exploring strategies to mitigate them, we aim to enhance the overall accuracy and reliability of our model.

The integration of the ML model into Roche's (DQO) exploratory tool described in Chapter 3 facilitated compliance with RBM principles and improved the accuracy of identifying risk and potential data quality concerns wherein data points required requerying.

This improvement has enabled the prompt initiation of preventative and corrective actions such as but not limited to: an evaluation of the risk and impact of requerying on CTQ data points versus non-critical data as well as improvements in query texts to mitigate unnecessary requerying. As the DQO tool shows a differentiation between functions and sites along with the number of reopened queries, users can also identify specific roles that could benefit from guidance or training to minimize the number of queries. These insights aid in streamlining workflows through risk detection as well as identifying pain points that can be categorically resolved through preventative and corrective actions and thereby enabling maintenance of data quality against the defined benchmarks.

CHAPTER 2. ANALYZING THE SENTIMENT OF QUERY RESPONSES

MATERIALS

In order to analyze the sentiment of query responses, assess whether the CRF field should be updated based on the query response text, and compare the presence of contradictory information in the CRF field Audit trail, we compiled a dataset of 100,000 queries randomly selected across various therapeutic areas in clinical trials. They were further

labeled by domain experts using a human labeling guide. The labeling was based on the implication of a data change in each query. To determine the label, the experts considered the query response, taking into account the query input text as context. Overall, there were three possible human labels assigned to each query:

- The query response implies a data change (“Data changed”).
- The query response implies the data was not changed (“Data not changed”).
- The query response is unclear as to an implication of a data change (“Unclear”).

This approach allowed us to accurately categorize the queries and analyze their implications regarding data changes.

In order to train and evaluate our model, we then divided the complete human-labeled query data into three separate datasets: training, development, and testing. The training and development datasets were allocated 75% of the total data, while the remaining 25% was set aside for testing.

To further refine our model during training, we divided the training dataset into two components. Approximately 75% of the training data was used for actual model training, while the remaining 25% was utilized for development and fine-tuning.

This division of data into multiple sets enabled us to effectively train, validate, and evaluate our model's performance at different stages, ensuring robustness and accuracy in our results.

METHODS

To address RQ 2, we conducted analysis into the query responses, grouped by the human label of an implication of a data change. Using this analysis, we engineered features to encode the insight in order to be used as input for two of the four XGBoost machine learning models.

Feature Engineering

As the first step of our exploratory data analysis (EDA) on the query responses, we derived several features to better understand and analyze the data. We conducted EDA by creating word clouds for each human label (Figure 3), which provided valuable insights into the patterns and characteristics of the query responses.

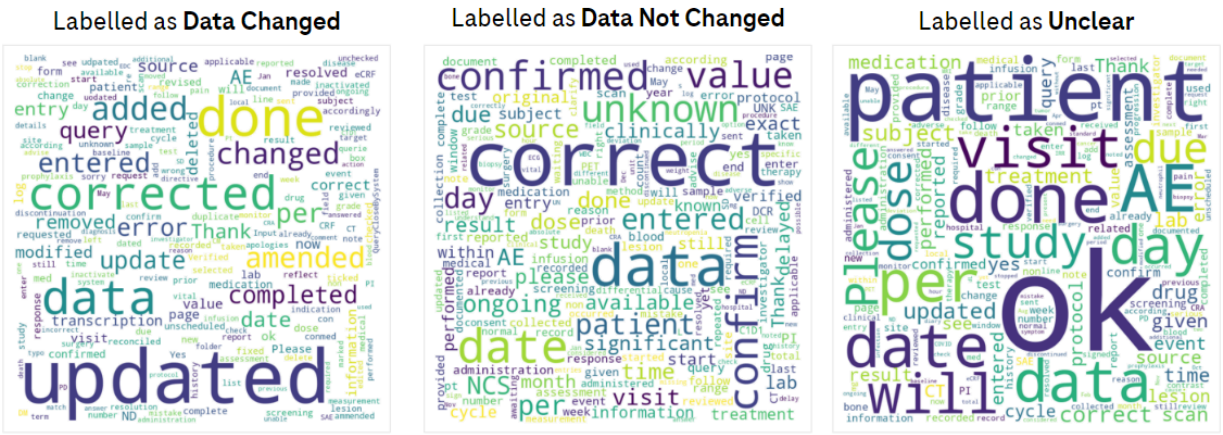

Figure 3. Word Clouds of Query Responses Grouped by Human Label, described in subsection “Materials”

In total, we derived 8 features from this analysis:

- The proportion of past tense verbs in the query response.
- The presence of a synonym for the word "updated" in the response.
- The presence of a synonym for the phrase "not updated" in the response.
- The presence of the word "not" in the response.
- The length of the query response in terms of the number of characters.
- The presence of numeric values in the response.
- The presence of the word "ok" in the response.
- The presence of the word "correct" in the response.

In addition to the engineered features, we expanded our feature set by employing the TF-IDF (Term Frequency - Inverse Document Frequency) technique (Equation 3) .

$$TF(t,d) = \frac{number\ of\ times\ t\ appears\ in\ d}{total\ number\ of\ terms\ in\ d}$$

$$IDF(t) = log(\frac{N}{1+df})$$

$$TF - IDF(t,d) = TF(t,d) * IDF(t)$$

Equation 3. TF-IDF (Term Frequency - Inverse Document Frequency) formulation, where t is a term within a document, d and N is the total number of documents.

The TF-IDF approach allowed us to identify important words and phrases based on their frequency and relevance within both the query text and responses data. To optimize performance, we determined that selecting the top 50 features from each (query text and responses) was most effective. Additionally, we utilized an n-gram [9] range of 1-4, encompassing both individual words as well as phrases of up to four words in length. This comprehensive approach enabled us to capture essential words and phrases from both query text and responses. This resulted in a robust set of features totaling 108.

These features were used in various combinations as inputs to the XGBoost model to further analyze their impact and contribution to the model's predictive accuracy.

Application of XGBoost Machine Learning Technique

XGBoost (eXtreme Gradient Boosting) is an open-source machine learning library that offers an efficient and scalable implementation of gradient boosted decision trees. The core algorithm is an enhancement over traditional gradient boosting techniques, addressing regularization through the addition of a penalty term to avoid overfitting (Figure 4). XGBoost supports regression, classification, ranking, and user-defined prediction tasks, and can be integrated with most of the data science frameworks [8].

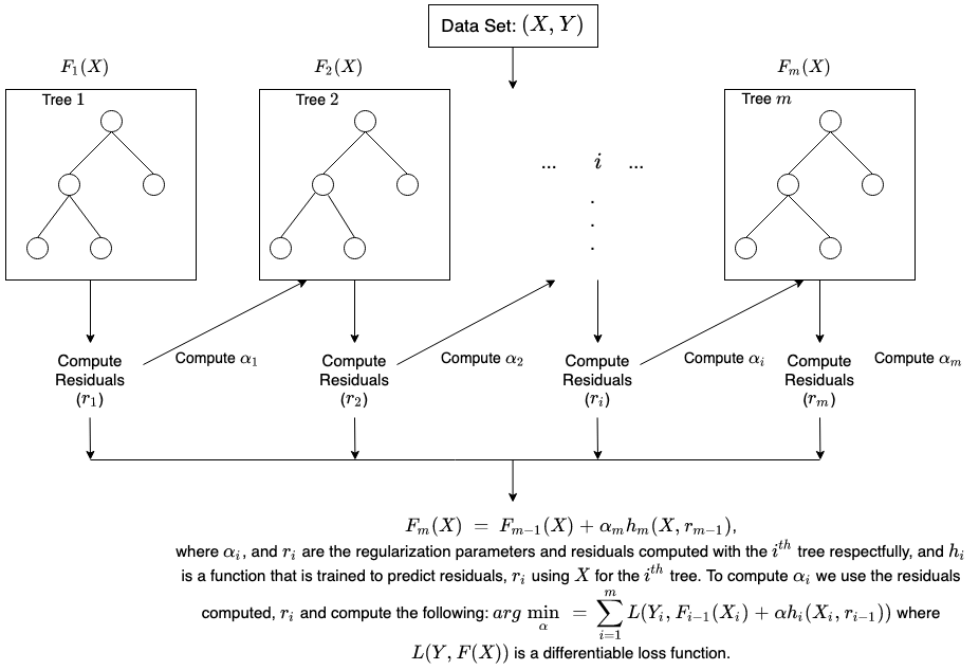

Figure 4. XGBoost Architecture Diagram [10]

We trained a total of four XGBoost-based models, each using different subsets of features. The four models trained were as follows:

- Model trained with 8 derived features: This model utilized the 8 features derived from the analysis, excluding the TF-IDF features (further referred as Model 1).
- Model trained with all features: This model incorporated all available features, including the 8 derived features as well as the TF-IDF features (further referred as Model 2).
- Model trained with TF-IDF features for the query responses and input text only: This model focused on utilizing the TF-IDF features specifically for the query responses and input text (further referred as Model 3).

- Model trained with TF-IDF features for the query responses only: This model utilized the TF-IDF features exclusively for the query responses (further referred as Model 4).

By training these models with different feature subsets, we aimed to explore the impact of various features on the performance of the models. The balanced accuracy score (Equation 2) served as a metric to assess the overall effectiveness of each model in terms of correctly identifying both true positives and false negatives. In addition to the balanced accuracy score, we leveraged the Shapley Additive explanations (SHAP) Python library to explain the selected best performing ML model. SHAP is a well accepted tool for explaining the output of machine learning models. It's grounded in cooperative game theory and offers insights into the contribution of each feature to a particular prediction made by a machine learning model [11].

RESULTS

The labels 0, 1 and 2 in this section refer to data not changed, data changed and unclear respectively.

MODEL 1

Training a model on the 8 engineered features exclusively, provides 84.5%, 81.8% and 84.5% balanced accuracy on the train, development and test sets respectively. Confusion matrices are represented in Figure 5.

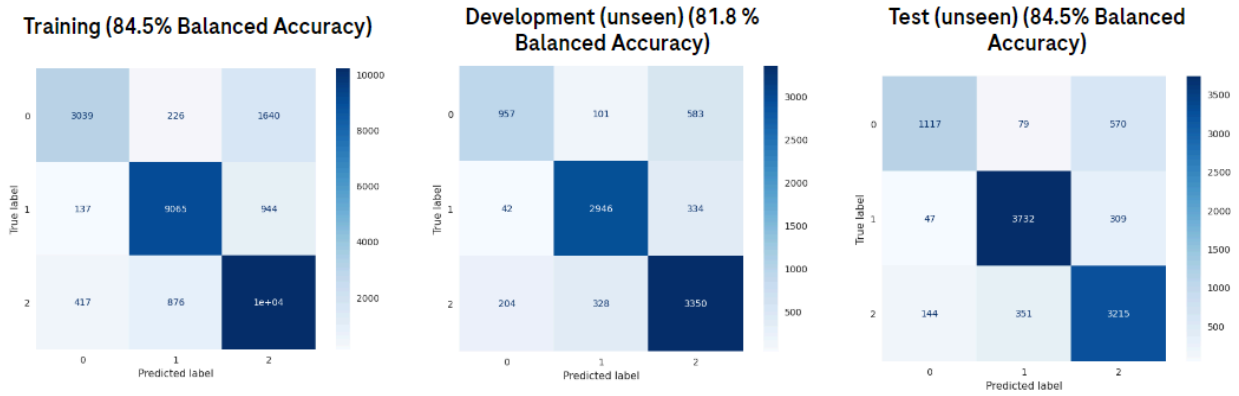

Figure 5. Confusion Matrices of XGBoost on Derived Features (Model 1). The predicted label of the model is shown along the X-axis which is compared to the ground truth on the Y-axis.

MODEL 2

Training a model on all of our input data provides 92.6%, 87.4% and 89.6% balanced accuracy on the train, development and test sets respectively. Confusion matrices are represented in Figure 6.

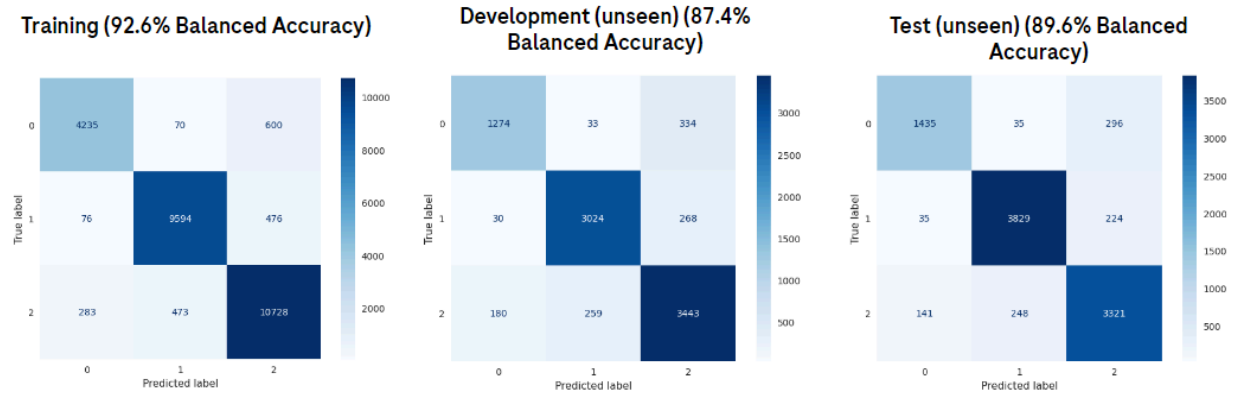

Figure 6. Confusion Matrices of XGBoost on All features (Model 2). The predicted label of the model is shown along the X-axis which is compared to the ground truth on the Y-axis.

MODEL 3

Training a model on only the TF-IDF vectors (query response & query text) provides 90.5%, 85.8% and 87.4% balanced accuracy on the train, development and test sets respectively. Confusion matrices are represented in Figure 7.

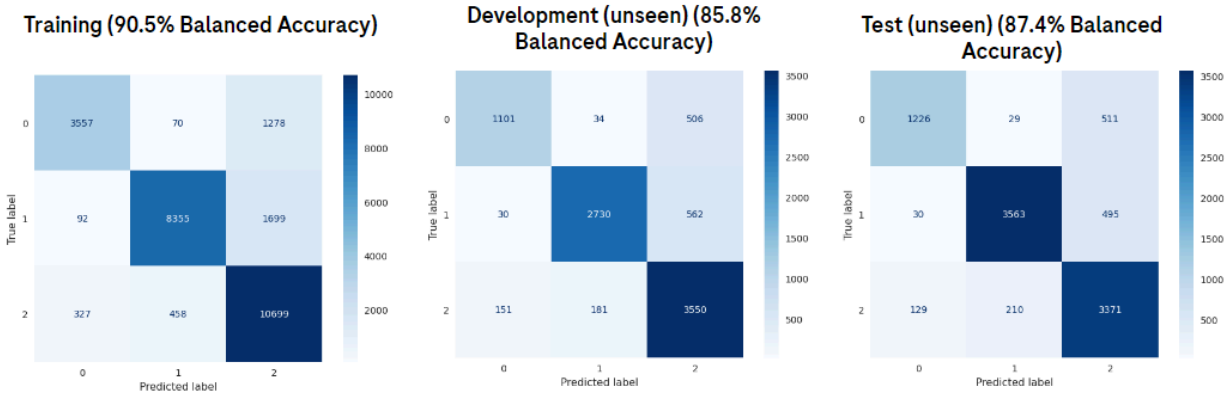

Figure 7. Confusion Matrices of XGBoost on All TF-IDF Features (Model 3). The predicted label of the model is shown along the X-axis which is compared to the ground truth on the Y-axis.

Model 4

Training a model on only the query response TF-IDF vectors provides 87.2%, 85.2% and 86.4% balanced accuracy on the train, development and test sets respectively. Confusion matrices are represented in Figure 8.

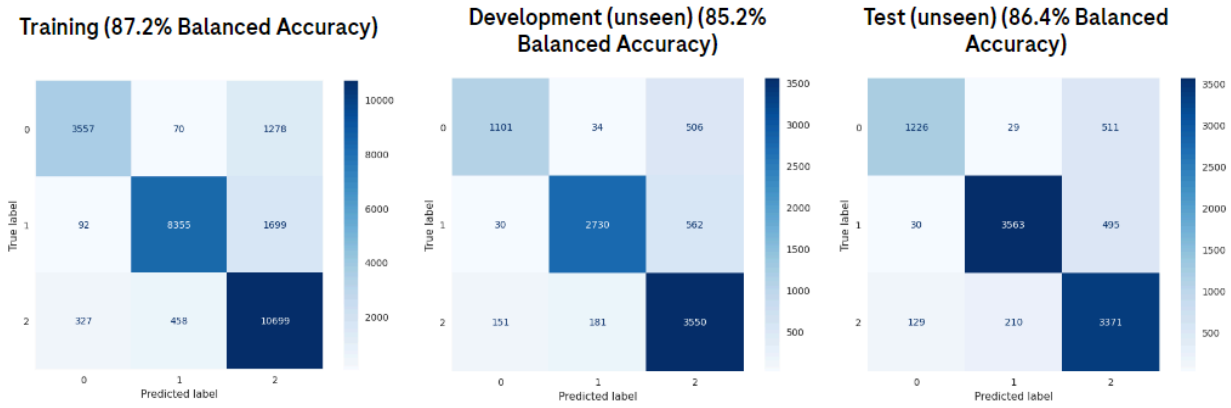

Figure 8. Confusion Matrices of XGBoost on Query Response TF-IDF Features (Model 4). The predicted label of the model is shown along the X-axis which is compared to the ground truth on the Y-axis.

Selected Best Performing Model

The final model chosen (model 2) displayed superior performance and was deemed the most effective among the models evaluated. To illustrate the significance of the features within this model, SHAP ML explainability was applied. This presents the feature importance as well as the impact of differing feature values on the model's prediction across the different classes (Figures 9, 10, 11). These visualizations highlight the model's strong logical explainability, further affirming its effectiveness and reliability.

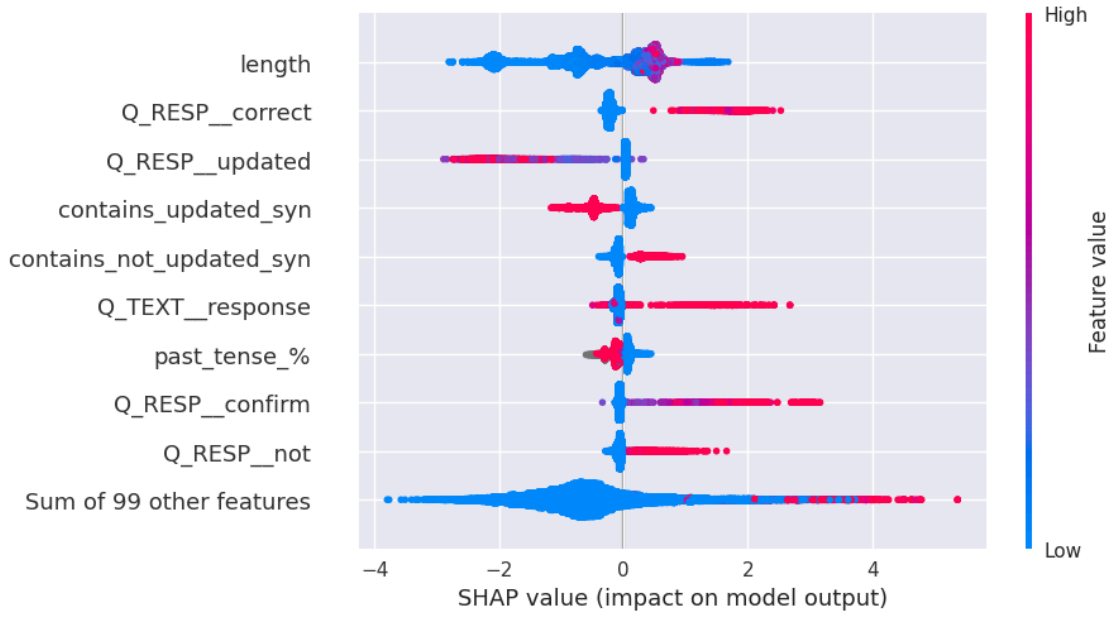

Figure 9. SHAP Plot of Feature Importance - **Predicting Not Leading to Data Changes.** This SHAP Beeswarm plot shows the feature importance of different feature values on predicting class 0 (not leading to data changes). In instances where the query response exhibits high TF-IDF values for terms such as "correct," "confirm," and "not," indicating their strong presence in the response, the model demonstrates a tendency to predict that these responses do not result in data changes. Similarly, when the TF-IDF values for the term "updated" are low, suggesting its limited presence in the query response, the model also leans towards predicting that these responses do not lead to data changes. Furthermore, the model has acquired the knowledge that if a query response contains a synonym for "updated" without an antonym, it is more likely to imply a data change.

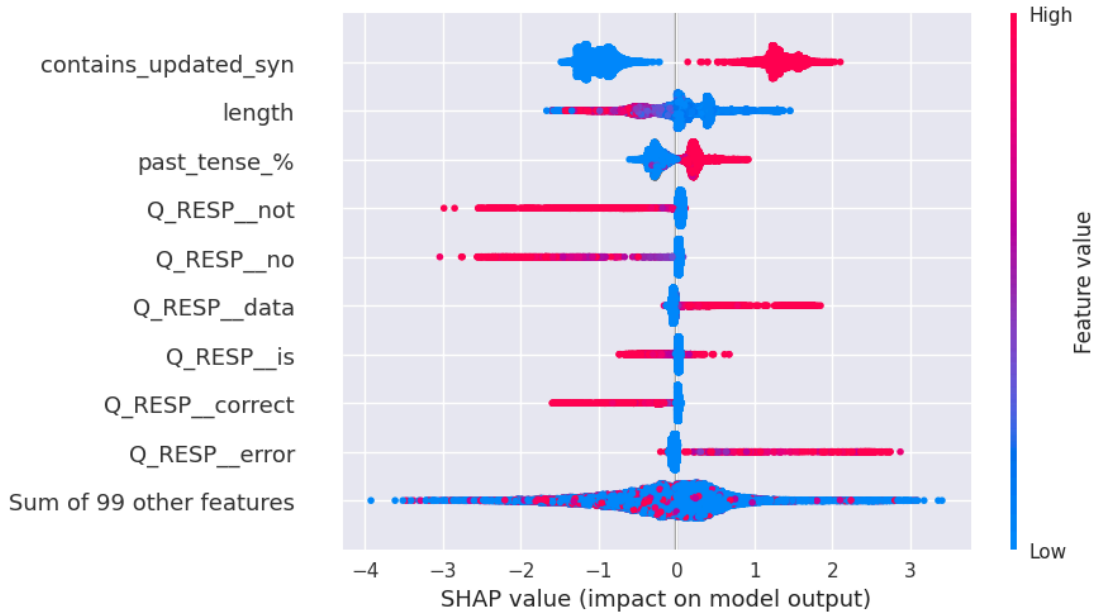

Figure 10. SHAP Plot of Feature Importance - **Predicting Leading to Data Changes.** This SHAP Beeswarm plot shows the feature importance of different feature values on predicting class 1 (leading to data changes). In the prediction of data changes, the presence of a synonym for "updated" in the response emerges as the most significant feature. In instances where the query response exhibits high TF-IDF values for terms such as "data" and "error," indicating their strong presence in the response, the model demonstrates a tendency to predict that these responses do not lead to data changes. Conversely, when the TF-IDF values for terms like "not," "no," and "correct" are low, suggesting their limited presence in the query response, the model has learned to slightly favor predictions indicating data changes, although their contribution is minimal.

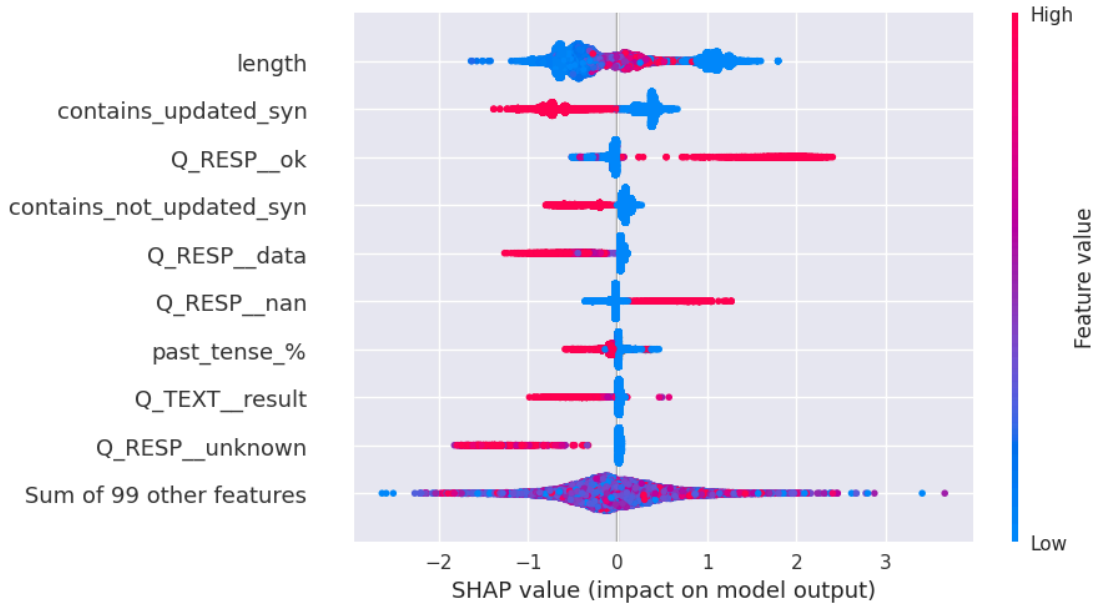

Figure 11. **SHAP Plot of Feature Importance - Predicting Unclear Response.** This SHAP Beeswarm plot shows the feature importance of different feature values on predicting class 2 (unclear). The model leans towards predicting "unclear" in two specific scenarios. Firstly, when the response is empty, and secondly, when the response does not contain a synonym for "updated" but does contain an antonym for "updated". These particular response patterns have been identified as having a correlation with "unclear" predictions.

DISCUSSION

Four models were trained and evaluated with the second model the best performing (trained on a 100 feature TF-IDF vector as well as 8 engineered features). Model 2 demonstrates effectiveness in determining if an eCRF data entry should be updated based on query and response text, revealing conflicting site activities and positively answering RQ 2.

The TF-IDF exclusive model with both the query text and response vectors outperformed the model exclusively trained on the query response vectors. Since the performance gain was only ~1% on the test data, we cannot conclude that the query text provides adequate context to the responses. Although the engineered features themselves have strong predictive power, they need to be paired with the query text and response TF-IDF vectors to maximize performance. After combining both, model performance improved by 5.1% on the test data (model 1 vs 2).

Due to the use of n-grams in the TF-IDF vectorizer, the vectors ultimately contained several stop words which may add unnecessary noise into the model. Having said this, the model does not overfit and from the explainability plots, the model does not appear to have learnt any irrational relationships with stop words.

Since the model is trained on human labeled data, human error may be present in the training data which can impact the quality of the model. Experts within the DQO team have conducted a qualitative evaluation of the model prediction within the tool and deem it to be high quality.

As with any ML model, there is the possibility of it exhibiting bias, including dataset used and methods applied. However, based on the extracted features from the TF-IDF vectors, the features appear to be very generic towards data change implications as opposed to specific to certain query attributes. Such as: CRF pages, CRF fields (with the exception of “AE”), countries, sites or patients.

The integration of the ML model into Roche's DQO exploratory tool described in Chapter 3 is particularly valuable in making sense of contradictory query responses and those in agreement with the audit trail in EDC system, thus increasing the accuracy of analytics that result in changes. Our integration of the ML model to DQO tool has demonstrated several specific benefits:

- It helps to identify instances of inattentive site performance.
- It helps to proactively mitigate potential data quality concerns for critical fields.
- It streamlines efforts for non-critical fields and queries that do not lead to changes.
- It improves the accuracy of the data changes visualization, based on the Audit trail data, by taking into account more queries that lead to changes in other fields.
- It provides insights into potential increases in the data quality backlog, enabling effective resource planning.

By leveraging ML-driven analytics, our research supports conscious risk-based decision-making by embedding RBM principles while also contributing to efficient resource allocation also aligning with RBM principles.

CHAPTER 3. PROPOSITION OF COMPREHENSIVE RBM ANALYTICS SOLUTION

The need for an integrated and all-encompassing RBM solution has been pervading the field of clinical trial development. Our investigation under RQ 3 revolves around the potential integration of the proposed models into a comprehensive and holistic RBM system that is consistent with the guiding principles of GCP. The overarching goal is to seamlessly integrate quality and risk management into clinical trial design and operation, thus preempting risks and recognising or avoiding potential issues at an early stage.

The solution we propose is based on three fundamental pillars, each fortifying the other to create a robust, end-to-end RBM suite capable of harmonizing processes across diverse clinical trial development activities.

1st Pillar: Novel AI-Driven Analytics

At the forefront of our solution lies the development and evolution of innovative ML models. These models, as demonstrated by the two previous ML models, are necessary components in this solution. Their workflow is instrumental for adapting to evolving use cases, which emphasizes their important role in the constantly changing RBM landscape.

2nd Pillar: Tools for Specific Areas

In the second pillar, the focus zooms onto domain-specific tools aimed at providing tailored analytics. The example here is Roche's DQO exploratory tool, released for over 450 studies. Integrated with ML models from the first pillar, this tool provides real-time insights and risk-based solutions, supporting identification of the data quality challenges.

The DQO tool offers a suite of functionalities encapsulated in tabs, each offering distinctive insights into data quality (here green color is used to emphasize the areas where the ML models are integrated):

1. Outstanding Queries: It provides an overview of all the current EDC queries in a study as well as trends in how long queries have been open across different groups and locations. Filters allow focusing on critical data points or specific forms.
   Insights: Spot trends in critical data queries or queries taking too long to answer. Identify sites or areas that need additional support with queries.
2. Leading to Changes: It shows how queries affect data by revealing the number of queries that cause changes in the EDC system. It supports grouping by different factors like marking group, site, or automated edit check.
   Insights: Identify queries that do not lead to any changes in data, indicating potential inefficiencies. Check if queries align with the study protocol.
   Our ML model predicting contradictory query responses is seamlessly integrated within this part of the DQO tool, providing enhanced oversight for maintaining the quality of critical fields in instances of contradictory responses.
3. Queries per eCRF Page/Question: It presents a comprehensive overview of Data Quality activities in relation to the eCRF, aiding in the identification of pain points.
   Insights: Identify specific forms that cause many queries or assess whether data cleaning complies with quality principles.
4. Queries Over Time: It shows the history of query activity in the study, helping to predict potential query backlogs. It tracks data cleaning activity and changes over time.
   Insights: Detects periods with less data cleaning or potential backlogs. Check how data cleaning relates to critical data.
5. Reopened Queries: It identifies which eCRF forms or questions often need queries to be raised multiple times. It can highlight confusion or unclear instructions in the queries.
   Insights: Detect patterns in queries being reopened and identify areas needing additional guidance or training for reducing re-queries.
   Our ML model helps predict reopened queries more accurately by integrating with existing data, regardless of how long ago similar queries were raised.

Currently, the DQO tool supports study teams in monitoring data quality activities. Here is an example of one of the testimonials from a user highlighting the applicability of RBM in the tool : "I have used it to show the team the queries age and also the patterns, similarities, association, dissociation between them, the sites and many other useful insights. We were able to target particular sites for a quick response turnaround and also we know when they are at their peak. We can clearly see when site resources are on holiday using patterns which show us the time/and days they take to respond to queries. With that info, we have changed our workload workflow." The tool, which is primarily

intended for data quality leads (DQLs), has also received positive qualitative feedback from other members of the RBM working group, supporting discussions on data cleaning and highlighting deviations from functional study documentation.

3rd Pillar: End-to-End RBM Suite

We suggest that this core pillar integrates all tools, starting from protocol design and culminating in interactive oversight, as offered by the DQO tool and others. Its development in an open-source fashion could provide significant benefits at an industry level. The suite aligns with QbD principles, emphasizing its main components:

1. Risk-Based Study Build System: Harmonizes study design by historical protocol comparisons, auto-generating CRFs based on program data, and suggesting a suite of key risk indicators (KRIs) and thresholds.
2. Risk-Based Study Monitoring System: Provides a central dashboard and issue management system, connecting risk detection to resolution and offering actionable insights.
3. Oversights Component: This component offers a comprehensive overview across multiple studies, highlighting successful approaches and recommending areas for improvement or necessary training. It acts as a reflective tool, showcasing successful practices while suggesting actionable areas to enhance performance and bridge skill gaps.

An open source approach to the development of these pillars could help to ensure collaboration and standardization across the pharmaceutical industry, aligns with regulatory expectations and promotes collaborative efforts towards a more efficient and robust RBM framework.

The proposed three-pillar framework is pivotal for advancing RBM practices across the clinical trial landscape. Open-source development of these pillars holds intrinsic value for pharmaceutical companies and regulatory bodies like the FDA.

From a pharmaceutical company perspective, open source development encourages collaborative innovation and enables companies to leverage collective expertise, reduce redundancy and accelerate the development of RBM practices. It provides a standardized approach that ensures interoperability, reduces costs and accelerates adoption.

From a regulatory perspective (e.g. FDA): A standardized, open-source RBM suite meets the FDA's goal of promoting innovation while ensuring patient safety and data integrity. It provides transparency, improves reproducibility of results and facilitates compliance with evolving regulatory guidelines.

The first and second pillars, which encapsulate novel AI-driven analytics and the domain-specific DQO tool, are well positioned for the initial open-source release. Their potential for collective collaboration and industry-wide adoption is substantial, laying the foundation for standardized RBM practices.

There are already well-established solutions in the second pillar, in particular the GSM package [12], and the PharmaVerse initiative [13]. These developments have established themselves as highly promising, well-established frameworks in the clinical reporting landscape. The GSM package is characterized by standardized risk-based monitoring, which focuses on the analysis of KRIs. Roche's PharmaVerse initiative promotes collaborative development and bridges the gap between pharmaceutical companies by fostering open source clinical reporting packages. Our DQO tool is based on the TEAL framework. These existing open-source solutions are leading by example, promoting improved collaboration and efficient reporting methods. While open source initiatives drive the first and second pillars, there are also several commercial solutions covering the third pillar that offer comprehensive monitoring capabilities but usually come with high investment requirements.

Our proposal is to collaboratively develop the third pillar in an open-source fashion – the End-to-End RBM Suite. Fragmentation and the lack of standardized RBM practices can be mitigated through joint initiatives that ensure seamless integration and compatibility of diverse clinical trial development activities.

The importance of this third pillar lies not just in its functionality but also in its adherence to QbD principles. This foundational approach ensures the system's robustness, reliability, and adaptability by aligning its development with the principles of GCP.

The proposal of a comprehensive RBM analytics solution encompassing the three pillars marks a paradigm shift in improving the efficiency, reliability and proactive nature of clinical trial monitoring. Its open-source nature invites collaboration, encourages innovation and most importantly, ensures patient-centricity and data integrity in clinical research.

CONCLUSION

In the fast evolving landscape of healthcare analytics, our initiative leverages unstructured free text to extract key insights to proactively improve data quality. We leverage state-of-the-art machine learning frameworks: XGBoost and pre-trained SBERT: These models demonstrate high accuracies of 89.6% and 90%, respectively, in detecting inconsistencies within textual EDC query data in order to maintain the quality of the data in a RBM fashion.

The integration of these models into the Data Quality Oversight tool across 450 studies made available real-time data insights and risk-based solutions. Users' widespread adoption and positive qualitative feedback highlighted the transformative potential of our approach.

The proposed three-pillar model marks the first step towards a comprehensive RBM suite. The integration of ML models, domain-specific tools and proposed end-to-end RBM suite presents an unparalleled opportunity to shape a collective future where clinical trial development thrives on unified, robust, and standardized RBM practices.

REFERENCES

1. Stansbury, Nicole, Brian Barnes, Amy Adams, Ruth Berlien, Danilo Branco, Debby Brown, Paula Butler et al. "Risk-based monitoring in clinical trials: increased adoption throughout 2020." Therapeutic Innovation & Regulatory Science 56, no. 3 (2022): 415-422.
2. Vora, Lalitkumar K., Amol D. Gholap, Keshava Jetha, Raghu Raj Singh Thakur, Hetvi K. Solanki, and Vivek P. Chavda. "Artificial intelligence in pharmaceutical technology and drug delivery design." Pharmaceutics 15, no. 7 (2023): 1916.
3. Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean. "Efficient estimation of word representations in vector space." arXiv preprint arXiv:1301.3781 (2013).
4. Pennington, Jeffrey, Richard Socher, and Christopher D. Manning. "Glove: Global vectors for word representation." In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), pp. 1532-1543. 2014.
5. Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. "Bert: Pre-training of deep bidirectional transformers for language understanding." arXiv preprint arXiv:1810.04805 (2018).
6. Reimers, Nils, and Iryna Gurevych. "Sentence-bert: Sentence embeddings using siamese bert-networks." arXiv preprint arXiv:1908.10084 (2019).
7. Quinlan, J. Ross. "Generating production rules from decision trees." In ijcai, vol. 87, pp. 304-307. 1987.
8. Chen, Tianqi, and Carlos Guestrin. "Xgboost: A scalable tree boosting system." In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, pp. 785-794. 2016.
9. Jurafsky, Daniel, and James H. Martin. "N-gram language models." Speech and language processing 23 (2018).
10. Mishra, Abhishek. Machine learning in the AWS cloud: Add intelligence to applications with Amazon Sagemaker and Amazon Rekognition. John Wiley & Sons, 2019.
11. Lundberg, Scott M., Gabriel Erion, Hugh Chen, Alex DeGrave, Jordan M. Prutkin, Bala Nair, Ronit Katz, Jonathan Himmelfarb, Nisha Bansal, and Su-In Lee. "From local explanations to global understanding with explainable AI for trees." Nature machine intelligence 2, no. 1 (2020): 56-67.
12. Wu G, Wildfire J, Roumaya M, Kosiba N, Sanders D, Childress S, Ge L, Wang Z, McLaughlin C, Dickens C, Gans M, Anderson J (2024). gsm: Good Statistical Monitoring. R package version 1.9.0, https://github.com/Gilead-BioStats/gsm.
13. "pharmaverse", accessed January 10, 2024, https://pharmaverse.org/

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the authors at:

Author Name: Lawrence Dennison-Hall
Company: Roche Products Ltd.
Work Phone: +44 7721 855751
Email: lawrence.dennison-hall@roche.com

Author Name: Mariia Lapaeva
Company: F. Hoffmann-La Roche Ltd
Email: mariia.lapaeva@roche.com