



Novel Analytics for Risk-Based Monitoring

Harnessing Natural Language Processing for Query Data to Enhance and Optimize Data Quality

Lawrence Dennison-Hall
Mariia Lapaeva

Agenda

Motivation and Objectives

RQ 1: Detection of Reopened Queries

- Materials and methods
- Results
- Limitations and Discussion

RQ 2: Analyzing the Sentiment of Query Response

- Materials and methods
- Results
- Limitations and Discussion

RQ3: Proposition of Comprehensive RBM Analytics Solution

Conclusion

Motivation and Objectives

Motivation

- 1 Most healthcare data is unstructured

Radiology
reports

EDC queries

Medical records

Protocol deviations

Adverse event reports

- 2 Unstructured **free-text** often **contains vital information**, posing challenges for traditional processing methods.
- 3 Novel ML methods could be used to **create new knowledge** and facilitate decision-making from text

Motivation

- ✓ Every EDC query incurs a time and resource cost
- ✓ The cost to resolve an EDC query is between \$28 and \$225 [\[1\]](#).
- ✓ Persistent querying can significantly increase the burden on sites and may be a signal of inefficiency or misunderstanding
- ✓ The need for a **comprehensive open-source RBM solution** is crucial to support cohesive processes across various clinical trial development activities.



Objectives

Provide a generalized RBM solution, able to **proactively detect data quality concerns** within previously unanalyzed textual EDC **query data** to enable insights that can help inform **risk-based decision making** across various clinical trial development activities.

- ❓ **RQ 1:** Can Machine Learning models effectively **identify reopened queries** not directly found within electronic Case Report Forms (eCRF) metadata?
- ❓ **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**?
- ❓ **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?



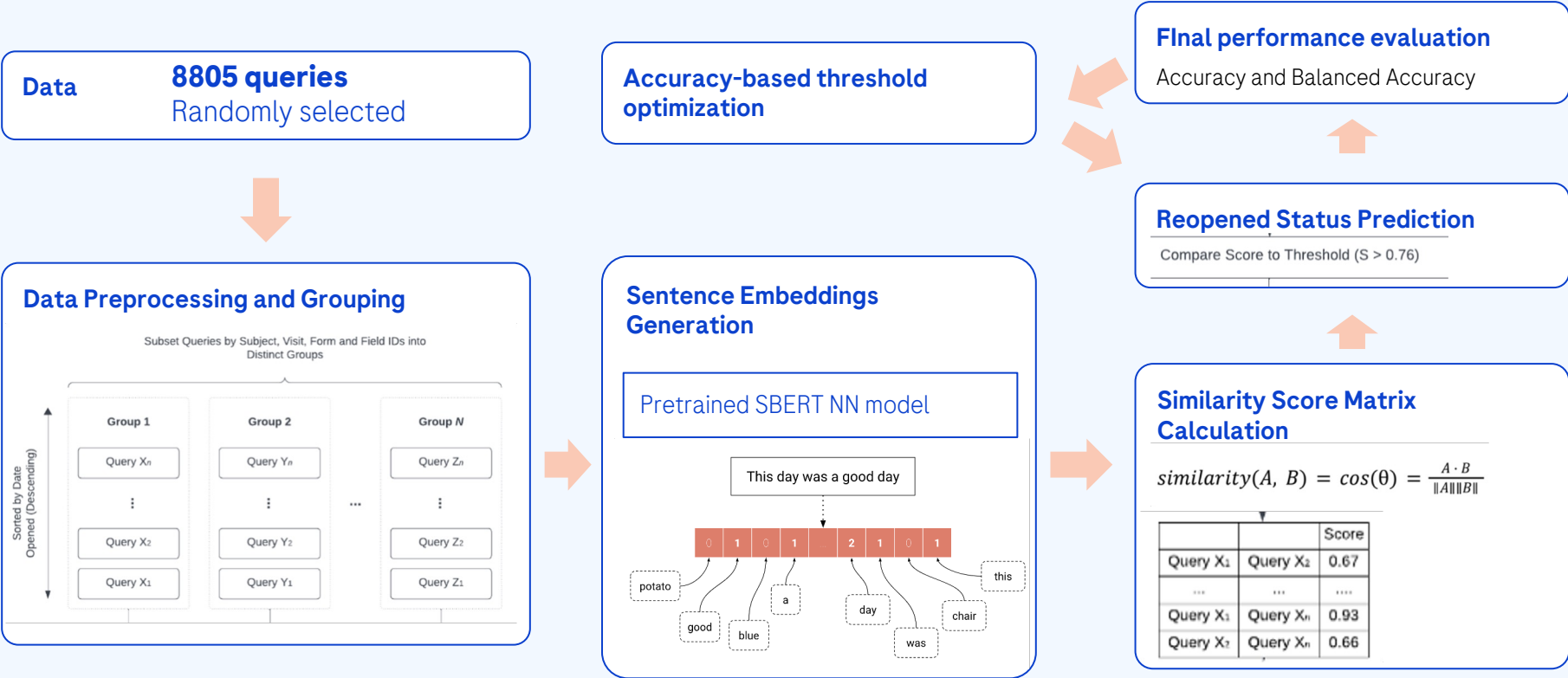
RQ 1: Can Machine Learning models effectively **identify reopened queries** not directly found within electronic Case Report Forms (eCRF) metadata?

Materials and methods

Results

Limitations and Discussion

RQ 1: Materials and methods



RQ 1: Results



Example true positive

Model successfully detected the similar sentences

Query 1:

Per protocol section, oral prednisone should be held on all days when pre-infusion methylprednisolone is given, however in this case the date of administration of oral prednisone is overlapping with the date of administration of premedication methylprednisolone. Kindly review and amend, else clarify.

Query 2:

Requery: Per protocol, oral prednisone is not expected to be administered on days when premedication with methylprednisolone is given. Kindly clarify if both drugs were administered on the same day, else update.

SUBJECT	VISIT	VISIT_CODE	FORM	FORM_LONG_NAME	FORM_ID	FIELD_LABEL	FIELD_NAME	score
0001	Week 1	W1	YYY	Premedications	00001	Date of Administration	XXX	0.785524070

RQ 1: Results

Accuracy is **90.5%**

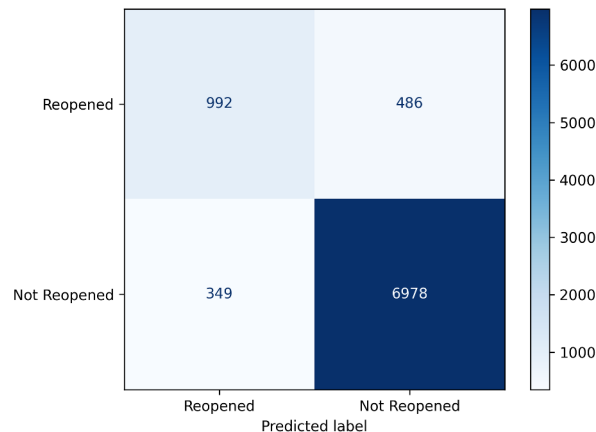
Balanced accuracy is 81.2%
(with similarity score threshold of 76%)



ML models could
effectively **identify
reopened queries**

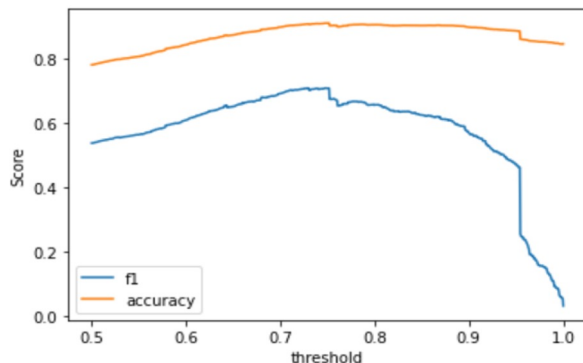
$$Accuracy = \frac{TP+TN}{TP+FN+TN+FP}$$

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



RQ 1: Limitations

- ① Chosen SBERT model, not pretrained on EDC query data.
- ① Threshold maximising accuracy on Roche query data may not be optimal for external query text styles



Discussion

- ML could detect reopen queries, some improvements could be done
- Facilitated compliance with RBM principles through the integration of the model into Roche's DQO exploratory tool
- Enables identification of preventative and corrective actions

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**?

Materials and methods

Results

Limitations and Discussion

Labels

100 000 queries
Randomly selected



Data changed

Data not changed

unclear

Feature Engineering

Derived features:

1. Past tense verb
2. Synonyms of “updated”
3. Synonyms of “not updated”
4. Presence of “not”
5. Response length
6. Numeric values
7. Presence of “ok”
8. Presence of “correct”

+ TF - IDF features

Data changed

Data not changed

unclear



RQ 2: Materials and methods

Data

100 000 queries
Randomly selected



Labels

Data changed

Data not changed

unclear



Feature Engineering

Derived features:

1. Past tense verb
2. Synonyms of "updated"
3. Synonyms of "not updated"
4. Presence of "not"
5. Response length
6. Numeric values
7. Presence of "ok"
8. Presence of "correct"

+ **TF - IDF features**



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

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

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

RQ 2: Materials and methods

Data

100 000 queries
Randomly selected



Labels

Data changed

Data not changed

unclear



Feature Engineering

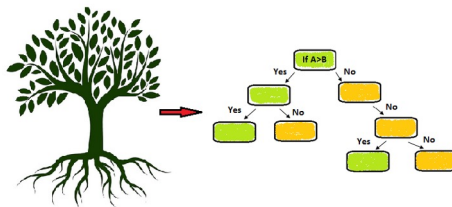
Derived features:

1. Past tense verb
2. Synonyms of “updated”
3. Synonyms of “not updated”
4. Presence of “not”
5. Response length
6. Numeric values
7. Presence of “ok”
8. Presence of “correct”

+ **TF - IDF features**



XGBoost (eXtreme Gradient Boosting) training



Final Performance evaluation and choosing the best model

Data changed	Data not changed	unclear
“Data is corrected”	“I confirm the data is correct”	“1”
“OK. Data has been updated”		“OK”

RQ 2. Results

Trained **4 models** in total with varying subsets of the available features

#	Training Data	Train	Dev.	Test
1	Engineered Features	84.5%	81.8%	84.5%
2	Engineered Features & TF-IDF Vectors	92.6%	87.4%	89.6%
3	TF-IDF Vectors Only	90.5%	85.8%	87.4%
4	Query Response TF-IDF Vectors Only	87.2%	85.2%	86.4%

Performance is provided in terms of balanced accuracy:

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



ML models could effectively
**determine query response
sentiment**

RQ 2. Results

Integration into Data Quality Oversight tool

ML prediction
of query response
sentiment

Data supposed to be:

Changed

Not changed

Unclear



Actual EDC Audit
trail data

Data actually was:

changed

Not changed

Unclear

=

Label in the DQO tool

Contradictory

Unclear Response

In agreement

RQ 2. Results

Query text	Query response	ML prediction (supposed to be)	Actual Audit trail EDC data	Label in DQO tool
"Please select appropriate Lab from dropdown list and respond to this query, else confirm if required lab is not present in drop down list."	"Data is updated"	Data changed	Data NOT changed	Contradictory
"Thank you for providing the response, however the data entered NA is non conformant. Please select ND - 'NOT DONE' from the drop down list if this is not applicable."	"U"	Unclear	Data changed	Unclear Response
"Date is not within the window of the Visit Date for the corresponding visit. Please verify your entries."	"confirmed within the window"	Data not changed	Data not changed	In agreement

← **Requires additional attention,** especially for **critical** data point

RQ 2. Limitations

- ① Feature choice
- ① Dataset and human labelling
- ① ML model

Discussion

- ML models could effectively **reveal contradictory activities, following RBM principles**
- The best performing was trained on a 100 feature TF-IDF vector as well as 8 engineered features

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?

Innovation challenge proposal

- 1st pillar
- 2nd pillar
- 3rd pillar

RQ 3. Innovation challenge proposal

Comprehensive RBM Analytics Solution

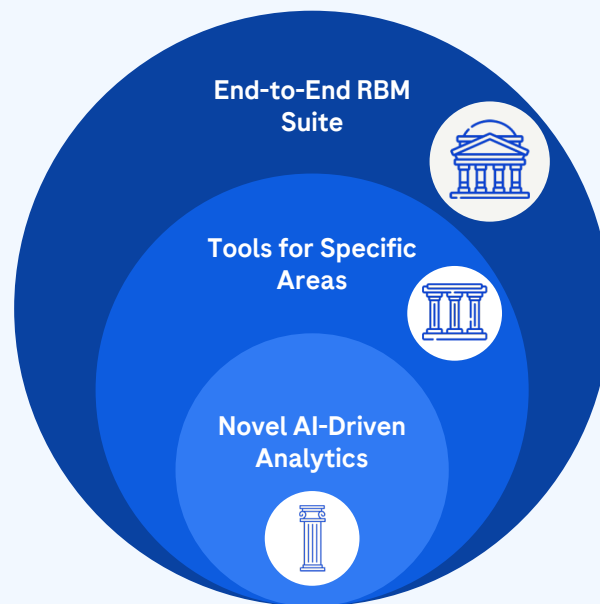


Goal: Seamlessly integrate quality and risk management into clinical trial design and operation, thus preempting risks and recognising or avoiding potential issues at an early stage



How: Integrate the proposed models into a comprehensive and holistic RBM system that is consistent with the guiding principles of GCP

Open Source



3 pillars model

RQ 3. Innovation challenge proposal - 1st pillar



Novel AI-Driven Analytics

- RBM landscape is constantly changing
- AI-driven analytics capabilities:
 - Discover new knowledge
 - Predict and highlight risks
 - Increase data accuracy
 - Analyse complex problems
- Our 2 proposed ML models are instrumental for the RBM landscape



Necessary!

Potential for Open-source



3 pillars model

RQ 3. Innovation challenge proposal - 2nd pillar



Tools for specific areas

- Domain-specific tools aimed at providing tailored analytics
- Example: **Roche's Data Quality Oversight (DQO) exploratory tool**, released for over 450 studies
 - Integrated with ML models from the first pillar
 - Tool provides real-time insights and risk-based solutions
- Currently, the DQO tool supports study teams in monitoring data quality activities

Potential for Open-source

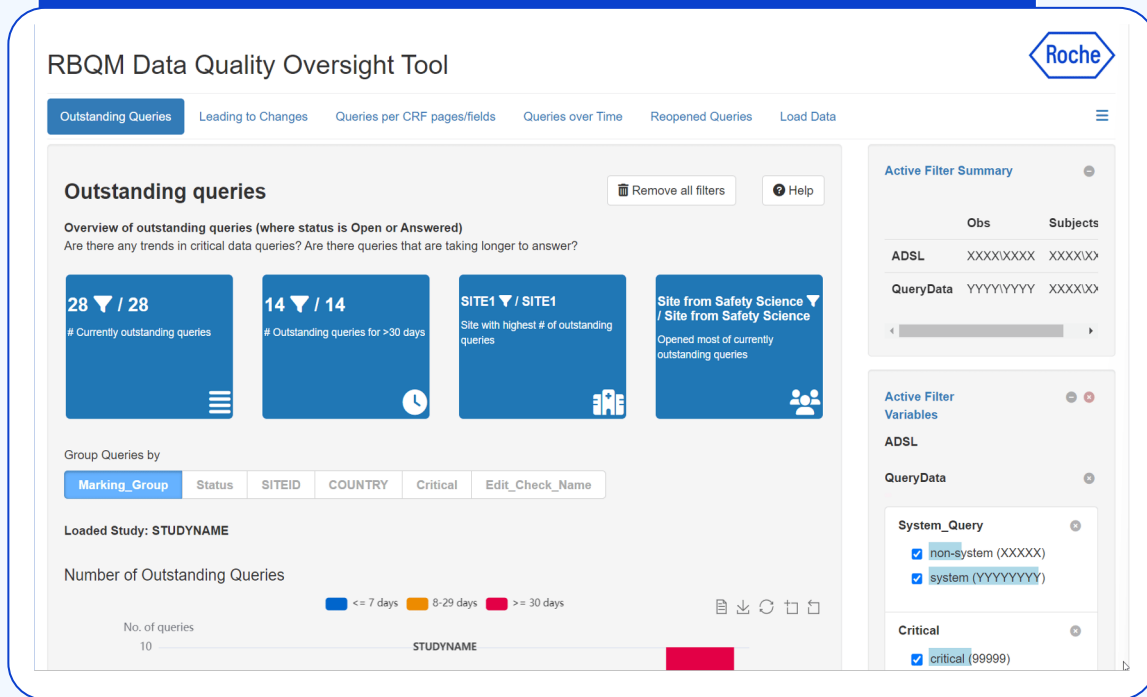


3 pillars model

Roche's Data Quality Oversight (DQO) exploratory tool

Outstanding Queries

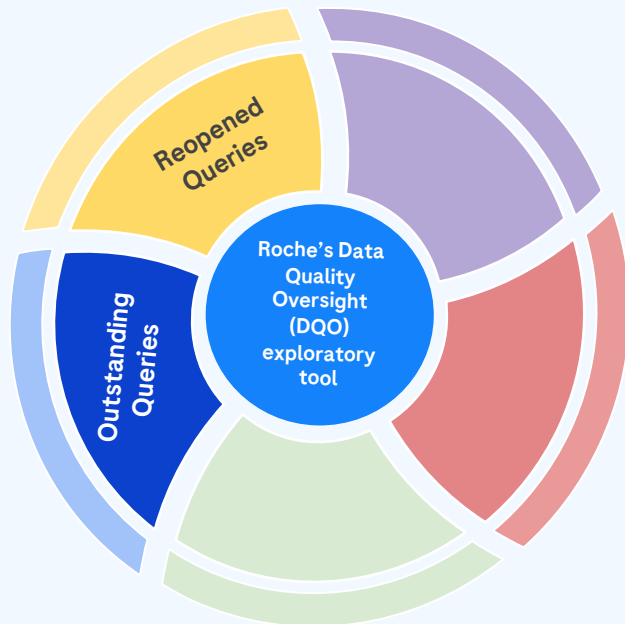
What is the current status of data cleaning activities on a study?
How many queries are open and for how long? Who raised the queries?
How many queries are answered and for how long?



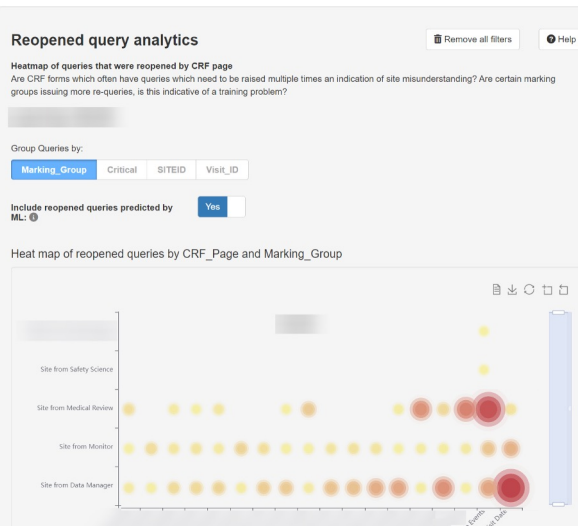
RQ 3. Innovation challenge proposal - 2nd pillar



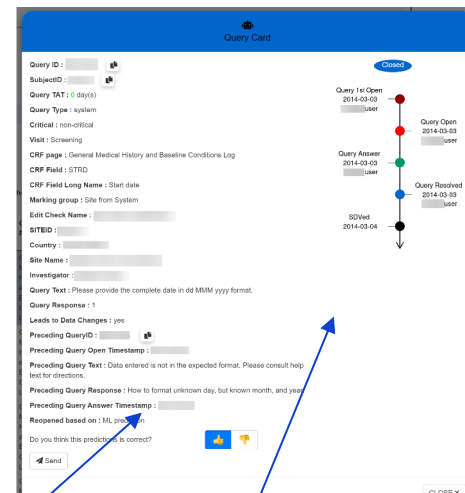
Tools for specific areas



How to identify problem pages and/or data points that require a high volume of re-queries, allowing for corrective actions to be made?



Preceding Query Text : Data entered is not in the expected format. Please consult help text for directions.
Preceding Query Response : How to format unknown day, but known month, and year.

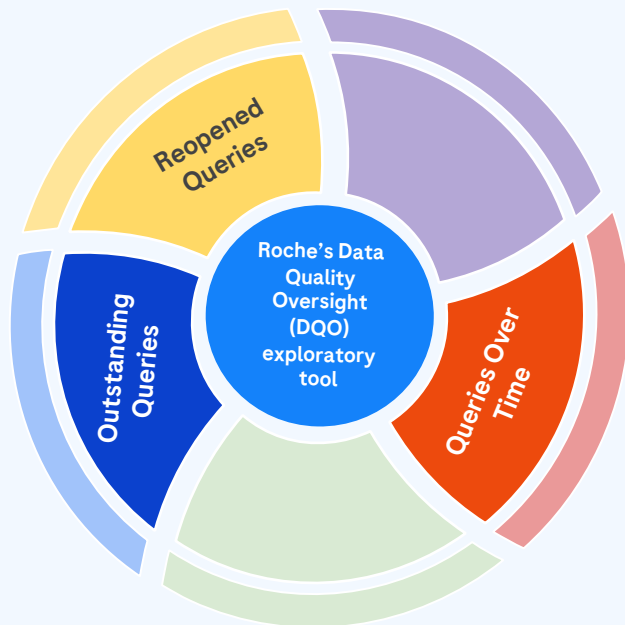


Query Text : Please provide the complete date in dd MMM yyyy format.
Query Response : 1
Leads to Data Changes : yes
Reopened based on : ML prediction

RQ 3. Innovation challenge proposal - 2nd pillar



Tools for specific areas



Are data being reviewed in a timely manner?

Do periods of low data cleaning activity increase the risk of future backlog

Number of queries over time

Remove all filters

Help

Query activity performed over the course of the study per month

Do periods of low data cleaning activity increase the risk of future backlogs? Has data cleaning activity changed in relation to critical and non-critical data? Is this expected?

Group Queries by:

None

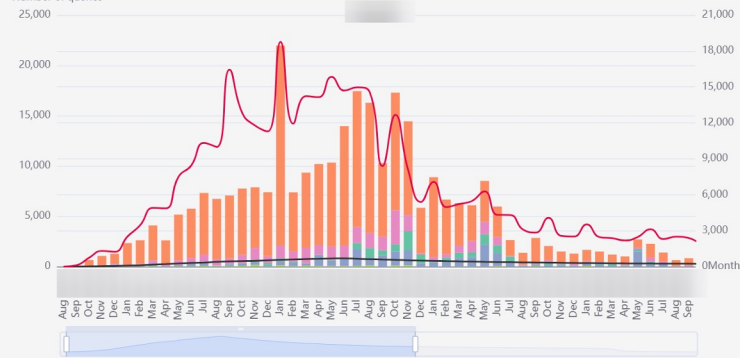
Marking_Group

CRF_Page

Number of Queries and Subjects on study per month

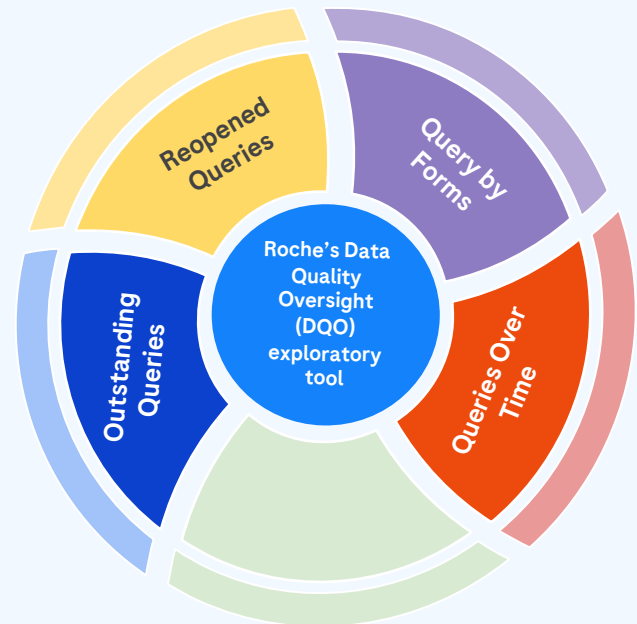
Subjects on study Entered CRF pages Monitor from Data Manager Site from 1/4

Number of queries

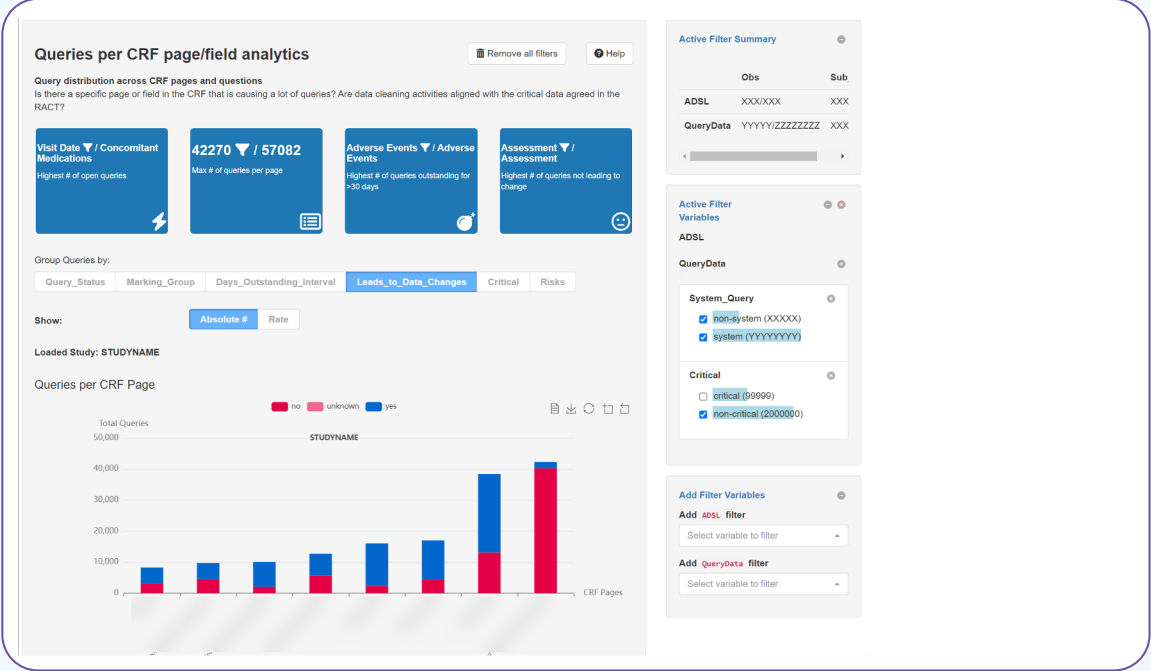


RQ 3. Innovation challenge proposal - 2nd pillar

Tools for specific areas

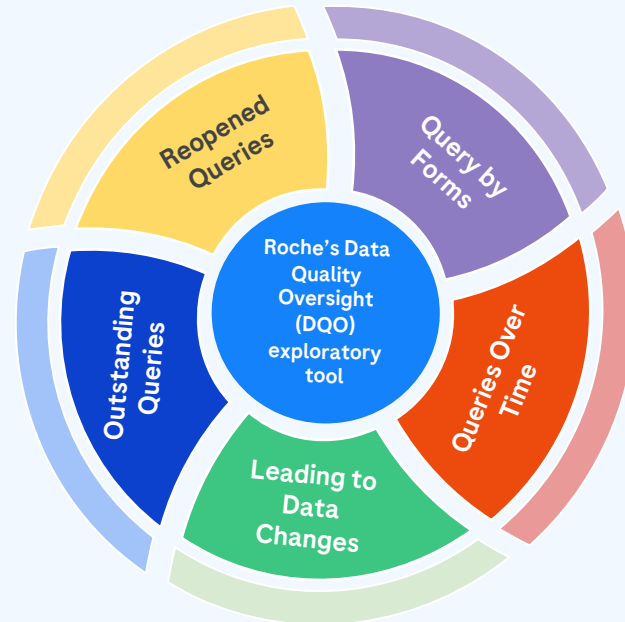


Which forms and fields have the highest query count? Is this expected?
 Are there any misfiring queries requiring an update?
 Is further guidance required via the CRF completion guidelines?

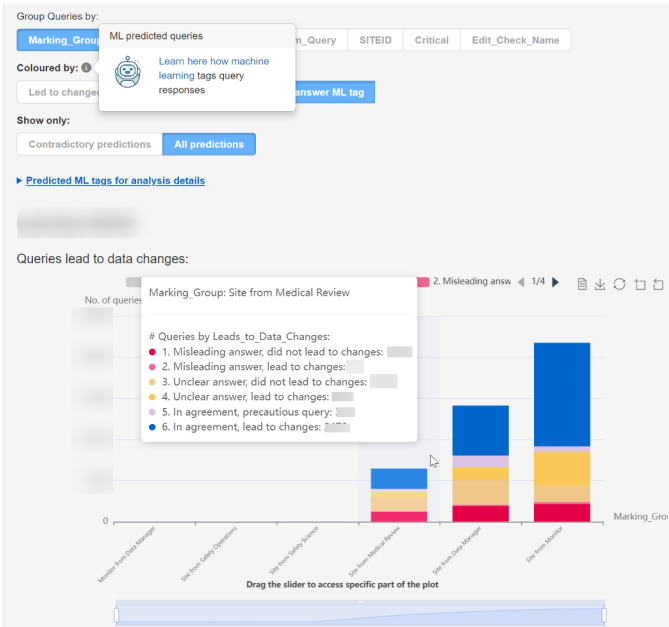
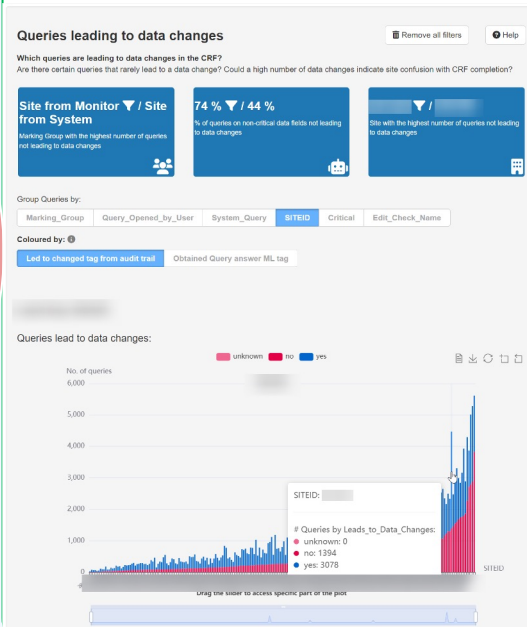


RQ 3. Innovation challenge proposal - 2nd pillar

Tools for specific areas



Focussing on critical-to-quality factors is important:
Are there certain queries that rarely lead to a data change? Could a high number of data changes indicate site confusion with CRF completion?



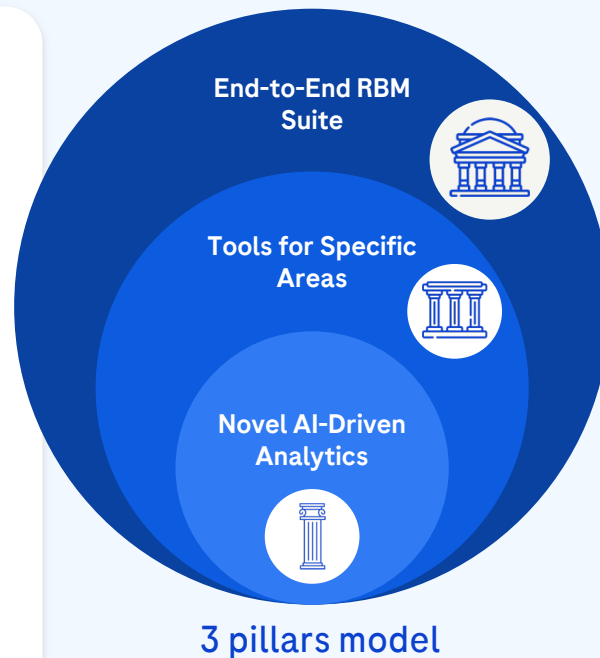
RQ 3. Innovation challenge proposal - 3rd pillar



End-to-End RBM Suite

Integration of RBM tools in an open-source fashion to provide industry level benefits:

- **Risk-Based Study Build System**
 - Harmonizes study design by historical protocol comparisons, auto-generating CRFs and suggesting key risk indicators (KRIs) and thresholds
- **Risk-Based Study Monitoring System**
 - Central dashboard for risk detection, issue management, and actionable insights.
- **Oversights Component**
 - Reflective tools
 - Overview across multiple studies
 - Showcase successful practices
 - Suggest actionable areas for enhance performance and bridge skill gaps



RQ 3. Innovation challenge proposal - 3rd pillar

End-to-End RBM Suite

Open source development:

- ✓ Encourages collaborative innovation
- ✓ Enables companies to leverage collective expertise
- ✓ Reduce redundancy
- ✓ Accelerate the development and adoption of RBM practices
- ✓ It provides transparency
- ✓ Improves reproducibility of results
- ✓ Facilitates compliance with evolving regulatory guidelines

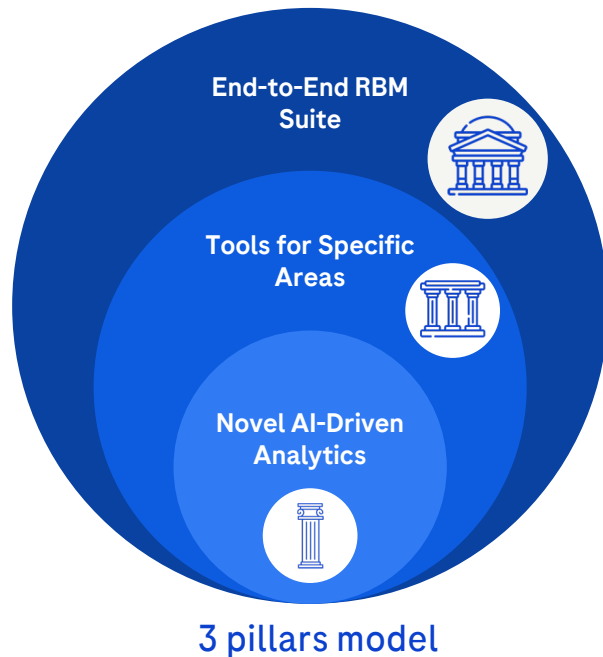


3 pillars model

Conclusion

Conclusion

- ✓ RQ1: ML models could effectively **identify reopened queries**
- ✓ RQ2: ML models could effectively **analyze query response sentiment**, uncovering contradictory activities in line with RBM principles.
- ✓ The integration of these models into the **Data Quality Oversight tool** across 450 studies made available real-time data insights and **risk-based solutions**.
- ✓ RQ3: Proposed **3-pillar RBM Solution** in open-source could: standardize practices, cut costs, and boost efficiency in clinical trials.





**Would you like the DQO tool
to be open-sourced?**



Join at menti.com | use code 1989 9812

Thank you!

Your comments and questions
are valued and encouraged:

lawrence.dennison-hall@roche.com

mariia.lapaeva@roche.com

February 27, 2024 | PHUSE US Connect 24

Doing now what patients need next