AI Assisted Tool for Regulatory Document Conformance Review – EIR Pilot Performance

Anna Hoffman[1], Daniel Priver[1], Ian Baker[1], Brian Campbell[2], Leslie Jackanicz[2], Jaqueline Guill[1], Indu Konduri[2]

[1]Booz Allen Hamilton

[2]Office of Regulatory Affairs, U.S. Food and Drug Administration

Abstract

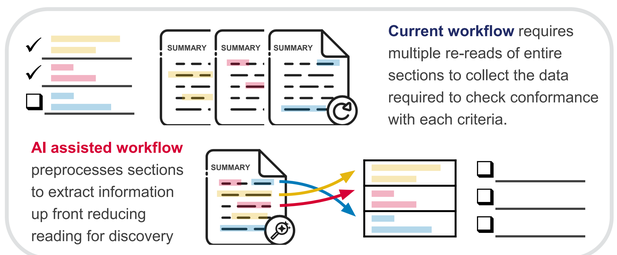
The power of LLMs in document review: Information extraction of foreign EIR IOM for criteria conformance

FDA employees review Establishment Inspection Reports (EIRs) every day to ensure adherence to established criteria. This project demonstrates the real-world benefit AI can provide in assisting conformance reviewers. For this Large Language Model (LLM) enabled pilot project we codified a subset of Investigation Operations Manual (IOM) criteria for EIRs. The tool extracts relevant text data from pdf EIRs for expedited conformance decisions by human reviewers. After review is complete, the tool automatically compiles reviewer comments into the desired feedback form. The pilot was delivered to DFHAFO users with a Jupyter notebook interface for testing.

Introduction

FDA sends investigators around the world to inspect establishments which produce food for American consumers. These inspections ensure U.S. food standards are being met, providing important protections for the health and safety of all Americans.

EIRs are an essential part of documenting inspectional findings by FDA investigators and must be reviewed to ensure adherence to established criteria prior to release. EIRs and corresponding IOM criteria are broken down by document section. Sections may be free text or tabular and can be pages long. The current review process is entirely manual.



Current workflow requires multiple re-reads of entire sections to collect the data required to check conformance with each criteria.

AI assisted workflow preprocesses sections to extract information up front reducing reading for discovery

Responsible AI **In concept planning we prioritized reducing burden on reviewers and minimizing added risk. This led to a process where AI makes no conformance decisions.**

FDA Reviewers will continue to make all conformance decision aided by AI enabled review. AI based information extraction and sorting completes most, if not all, of discovery work for reviewers. The extracted and sorted information increases reviewers' confidence in their decisions as documents are processed and when performing final review of completed work.

Materials and Methods

Data

7000 EIRs from FY18-23

219 FY23 IOM Criteria

SME Interviews on 19 criteria

Tools

python, jupyter, aws, Hugging Face

Development

Prepared EIR data for development through data ingestion pipeline creation, data cleaning, and developing parsing methods. This work was the basis of the document ETL pipeline within the tool.

At the same time, initial development of background LLM integration, free text processing algorithms, NLP search functions, and semantic search pipelines began.

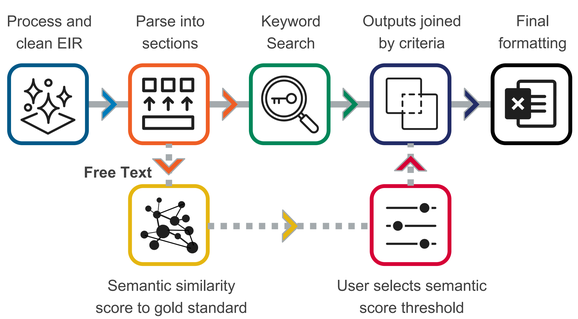
Interviews were conducted with reviewers serving as criteria and EIR SMEs to establish intent, keywords, nuances, acceptability standards, and required inclusions of each individual criteria.

The findings from SME interviews were incorporated with base language processing functions to create custom algorithms which searched text for keywords and keyword relationships for each criteria. Algorithms were developed on a training set of EIRs.

Once algorithms were complete, a random set of 500 EIRs were selected to create the gold standard dataset for each semantic search criteria. The keyword results from these EIRs for free text based criteria were embedded to vector format and saved for comparison against future EIRs.

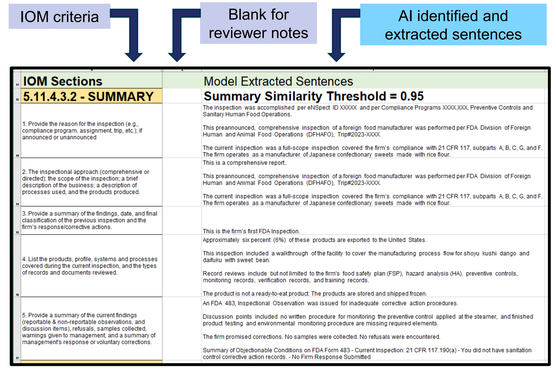
Tool Workflow

The tool takes EIRs and evaluates data by individual sentences to extract all sentences relevant to each criteria. Outputs are formatted to integrate into existing office workflow where reviewers document notes in a template Excel.



Output

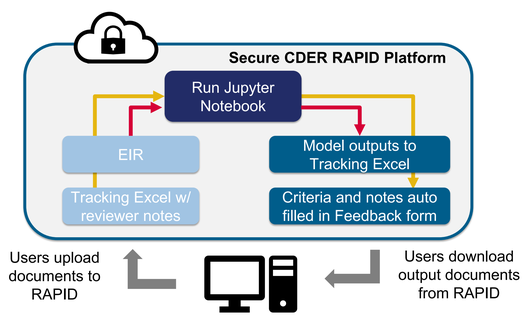
Integration with existing workflow was prioritized to reduce change management. Reviewers take notes in an Excel that lists out criteria, we have the model outputting extracted sentences in line with these criteria for easy review.



Pilot

To provide access to users and allow for user testing, the tool was built into a pilot with a Jupyter notebook interface backed by a python repository, the embedded gold standard datasets, and blank template documents. The pilot is hosted on a CPU AWS node with access granted by reviewer team leadership on an individual basis.

In addition to the AI assisted review tool, we built a BPA tool which processes all notes made by reviewers during their review. When users upload their completed review it compiles, formats, and outputs all comments into a standard feedback form.



Data Security **Users load a documents into memory for processing. Once outputs are created, refreshing the notebook clears all data from memory.**

Results and Discussion

The tool was evaluated against a hand annotated corpus of 100 EIRs. Performance metrics are shown below for selected individual criteria algorithms.

Criteria	Model	Accuracy	Precision	Recall	F1 Score
Summary 1	Semantic Search	93%	97%	97%	.97
Summary 2	Semantic Search	92%	92%	100%	.96
Summary 3	Semantic Search	92%	92%	100%	.96
Summary 4	Semantic Search	90%	96%	93%	.94
Summary 5	Semantic Search	93%	100%	93%	.96
Admin 1	Rule-Based NLP	92%	94%	95%	.94
Admin 2	Rule-Based NLP	92%	98%	93%	.96
Admin 3	Rule-Based NLP	95%	95%	100%	.97
Admin 5	Dependency Parsing	90%	91%	90%	.90
Admin Report 1	Rule-Based NLP	95%	100%	94%	.97
Admin Report 2	Rule-Based NLP	95%	96%	100%	.97
Admin Report 3	Rule-Based NLP	88%	100%	87%	.93

The tool performed well for all criteria algorithms. Rule-base NLP algorithms utilized regular expression (regex) and generally scored highest as these algorithms evaluated data contained in standardized tables with known structure and bounds.

AI Transparency **All algorithms are documented in common English in a technical briefing to ensure users understand how sentences are selected.**

Tool performance was strongly assisted by a recent update of the EIR template for FY23 from a fully free text document to a table-based form which populates standardized tables based on information contained in each EIR. The tool was built around this table-based format and any changes will require updates to the model.

Conclusion

Reviewers were able to test the tool on their own and provide anonymous feedback which was overall positive.

> *"I think it has excellent potential to increase efficiency in EIR review, particularly for some of the basic report data that has to be included in a report, and would allow us to focus on other aspects of the report that perhaps an AI tool wouldn't be a substitute for a review by a human"*

At this point only 19 of the 200+ IOM and Compliance Program criteria have been coded and evaluated. Next steps for this work is to expand to criteria in sections outside of Summary and Administrative Data. Development of new criteria will be completed at a faster pace now that the initial pipeline and infrastructure is complete.

With increased confidence in the model and advances in AI technology, there could be a future iteration of this model which provides conformance recommendations.

Acknowledgements: The team would like to thank the FDA ORA DFHAFO reviewer team for their support and indispensable SME assistance during development, and FDA CDER RAPID for the platform infrastructure and tool management that made this work possible.