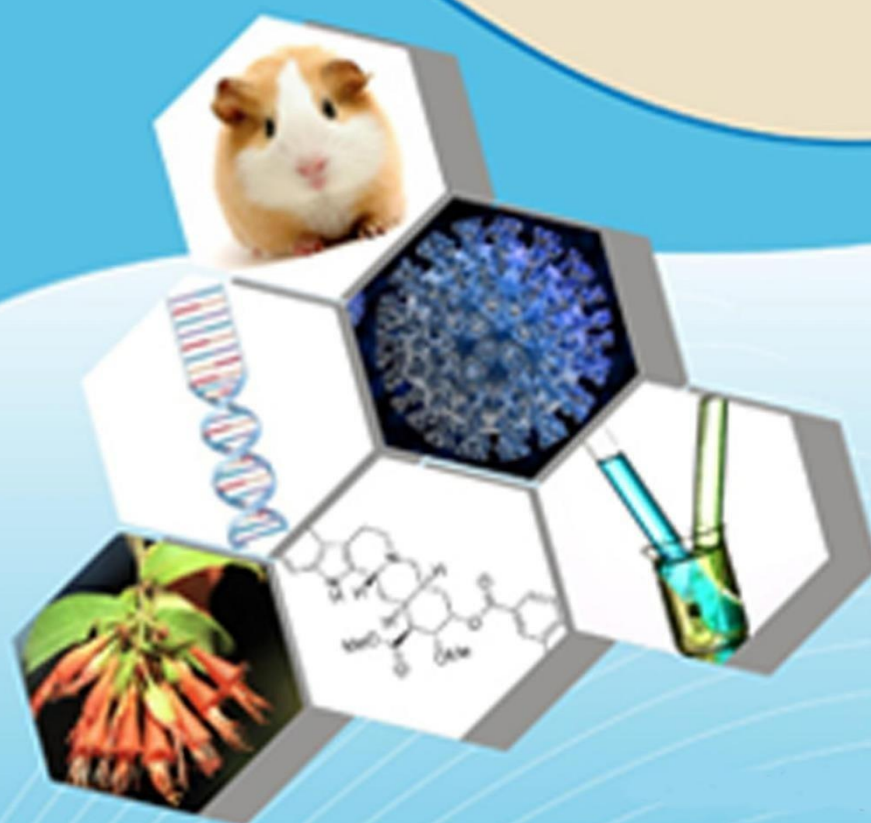




ISSN : 2347-2251
**Indo-American Journal of
Pharma and Bio Sciences**



www.iajpb.com

iajpb.editor@gmail.com
editor@iajpb.com



A Regulatory NLP Framework for Evaluating Data Sufficiency in FDA Meeting Request Submissions

¹Deepaben Bhavsar, ²Lynda Haixia Lin

¹Northeastern University- Boston, MA, USA

²Shanghai Institute of Pharmaceutical Industry, Shanghai, China

Abstract

During drug development, effective engagement with the U.S. Food and Drug Administration (FDA) is essential for achieving regulatory success. The process begins with submitting well-prepared FDA meeting requests, supported by clear, concise, and complete documentation that justifies the need for the interaction and facilitates a productive exchange of information. Nevertheless, the current process for evaluating the adequacy of information submitted to justify a meeting remains largely manual, time-intensive, and subject to interpretation. Amid the recent progress in the field of Natural Language Processing (NLP), a body of opportunity is emerging to create regulatory-oriented AI systems that automate the process of data adequacy in such submissions. The current paper suggests a regulatory NLP framework that should examine the completeness, clarity, and suitability of data in documents submitted for the FDA meeting requests. The article delves into the elements, approaches, and practical issues of implementation of such frameworks and how the same can help simplify regulatory interactions and maintain compliance with FDA expectations.

Keywords: Natural Language Processing, Regulatory Submissions, FDA Meeting Requests, Data Sufficiency, AI in Drug Development

I. Introduction

The increasing volume of regulatory submissions and the need for early engagement with the FDA have underscored the importance of timely and well-organized meeting requests. Meetings such as Pre-IND, Type B, End-of-Phase, and others provide sponsors with valuable opportunities to obtain FDA feedback on development strategies, clinical trial designs, and specific regulatory concerns. Nonetheless, regulatory bodies tend to focus on how sufficient data, context, and clarity are in such submissions. With a lack of sufficient information, meeting requests are either refused or postponed, leading to delays in the development schedule and resource misallocation. This challenge highlights the potential of Natural Language Processing (NLP) as a tool to refine, assess, and ensure the completeness of these documents. One of the branches of AI, called NLP, has demonstrated promising properties in regulatory settings regarding the classification of content in the text, sentiment analysis, and completeness verification of written documentation [1][2]. Although still in its early stages, this approach holds significant potential for application in FDA meeting request submissions. Variability in sponsor submissions, such as unclear clinical justifications, inconsistent formatting, and differing levels of detail, often leads to interpretive inconsistencies in regulatory review. By employing a regulatory NLP framework, these documents can be standardized, analyzed, and



scored in real time to determine whether they meet the FDA's baseline expectations for data adequacy [3][4]. The structure and content requirements of an FDA meeting request must be analyzed to place this proposal in the proper perspective. These generally consist of background about the product, the objectives of the meeting, a list of questions that the FDA reviewers need to investigate, and a briefing package comprised of supporting data. The assessment of adequacy in these areas depends on understanding regulatory guidance, precedent, and scientific context, factors that traditional rule-based systems cannot fully capture. However, NLP models trained on extensive datasets of previous submissions and FDA responses can effectively generate such contextual insights [5][6]. The paper will present a regulatory framework based on NLP, which will have the ability to evaluate the semantic and structural integrity of meeting request documents. The framework is created to identify potential critical aspects, assess the situation, and give a risk rating according to the FDA acceptability levels.

2. FDA Meeting Request Submissions: Background and Problems

It is important to recognize that the regulatory framework governing FDA meeting requests is defined by specific regulations and guidance documents that influence how such requests are prepared and submitted. The FDA offers several meeting types: Type A, Type B, and Type C, etc., each distinguished by its timing, purpose, and associated data requirements. As an example, a Pre-IND meeting (Type B) involves an elaborate blueprint in which nonclinical data, proposed clinical trial protocols, and specific regulatory questions are contained. Type C meetings, on the other hand, can be for a less broad topic and yet still require the precise description of the problem under discussion and the reasons behind it. Although there are formal instructional guidelines regarding submission requirements, sponsors frequently have difficulties in interpreting the extent and the type of data that is considered sufficient, resulting in unequal quality in submissions [7][8].

A key aspect of this challenge stems from the unstructured and inconsistently formatted nature of submission content. Applicants often provide highly narrative documents, make limited use of data visualization, or exclude essential contextual details, all of which impede efficient review. Differences in document structure, terminology, and completeness further increase the complexity of regulatory assessment. As a result, FDA reviewers must manually examine extensive submissions, cross-reference them with historical data, and determine whether the presented content is adequate for further evaluation. This process is time-consuming, labor-intensive, and introduces subjectivity into what is intended to be a standardized review mechanism [9]. Furthermore, the risk of inadequate data presentation is not determined solely by document length or word count. In many instances, ambiguity, lack of coherence, or insufficient contextual information can significantly diminish the overall utility of the submission. NLP provides a chance to deconstruct these problems systematically by assessing document semantics, structure, and language patterns in accordance with the expectations of the regulations. Although regular content validation programs can be used to check the formatting or the presence of a document, they cannot understand the flow of narration,



question congruency to background knowledge, or whether there are any gaps in data based on the historical standards of submissions [10][11].

The application of NLP is not directly a matter of automation but rather an improvement in the quality of the decision and submission readiness. Through mapping linguistic indicators, semantics at the sentence level, and comparison with a training set of approved and rejected meeting requests, an NLP framework is able to spot shortcomings that otherwise would not have been identified until the review step. This may allow sponsors to polish their contents prior to submission, lead to a successful regulatory friction, and increase the probability of getting timely engagement with the FDA [12][13].

3. Designing an NLP Framework for Regulatory Evaluation

Based on the identified challenges, the proposed NLP model will evaluate the adequacy of data within FDA meeting requests through a three-phase approach encompassing extraction, assessment, and scoring as shown in Figure 1. Basically, the system is required to analyze the unformatted text, identify the sections of the text that are relevant to the text, as well as to check the preparedness and mark their preparedness to be submitted to the FDA. This kind of scoring is not presented instead of human review, but it is, on the contrary, a supplementary aspect of human review, where it constitutes a standardized and evidence-based quality check.

This framework leverages Named Entity Recognition (NER) models trained on regulatory vocabularies to identify and label key content segments, such as product background, investigation plan, proposed indication, study design, and regulatory questions during the extraction phase. NLP algorithms are then applied in mapping these blocks on the specified FDA meeting request templates in order to identify missing or misaligned portions [14]. Dependency analysis and document segmentation systems enable the system to identify hierarchical relationships within the document, e.g., linking specific questions with their supporting justifications or underlying background rationales. Upon extraction, the evaluation engine utilizes the models depending on the transformer (e.g., Bidirectional Encoder Representations from Transformers (BERT) or variations) to rank the depth of context of each section. An example is that it proves a hypothetical question of dosing using pharmacokinetic information, or whether the foundations of a study structure provide quotations of correct nonclinical results. The step employs semantic similarity, relevancy scoring, and contrastive learning techniques to compare each document section against known sufficient posts and FDA responses, using transcripts from public meetings as the reference data source. This approach helps evaluate how closely the content aligns with established examples and regulatory feedback to ensure completeness and relevance in submissions [15][16].

The final phase involves composite scoring, where each content area is evaluated against predefined scales for clarity, completeness, and relevance, using historical submission data as a reference. The framework further determines a cumulative sufficiency score and interpretability characteristics that can either tell what gaps exist or what they have not clarified effectively. This will assist the sponsors to take action in relation to the issue without



necessarily waiting for the response of the regulatory authorities [17]. Most importantly, the framework is made flexible. In the event of such changing standards, the NLP models may be retrained to incorporate such standards, or rather, the models may be narrowed to incorporate them whenever new FDA guidelines are issued or new forms of submissions are used. In addition, explainability layers will make sure that sponsors can see the reason why certain features have attracted their attention, and it will be the stimulus to trust and openness in AI-based regulation inspection [18][19]. In the following section, the way such a structure could be practically applied to the working process of the sponsors, which includes the writing and quality control up to the submission feedback analysis and the post-submission feedback analysis, will be explained.

NLP_Framework_Pipeline

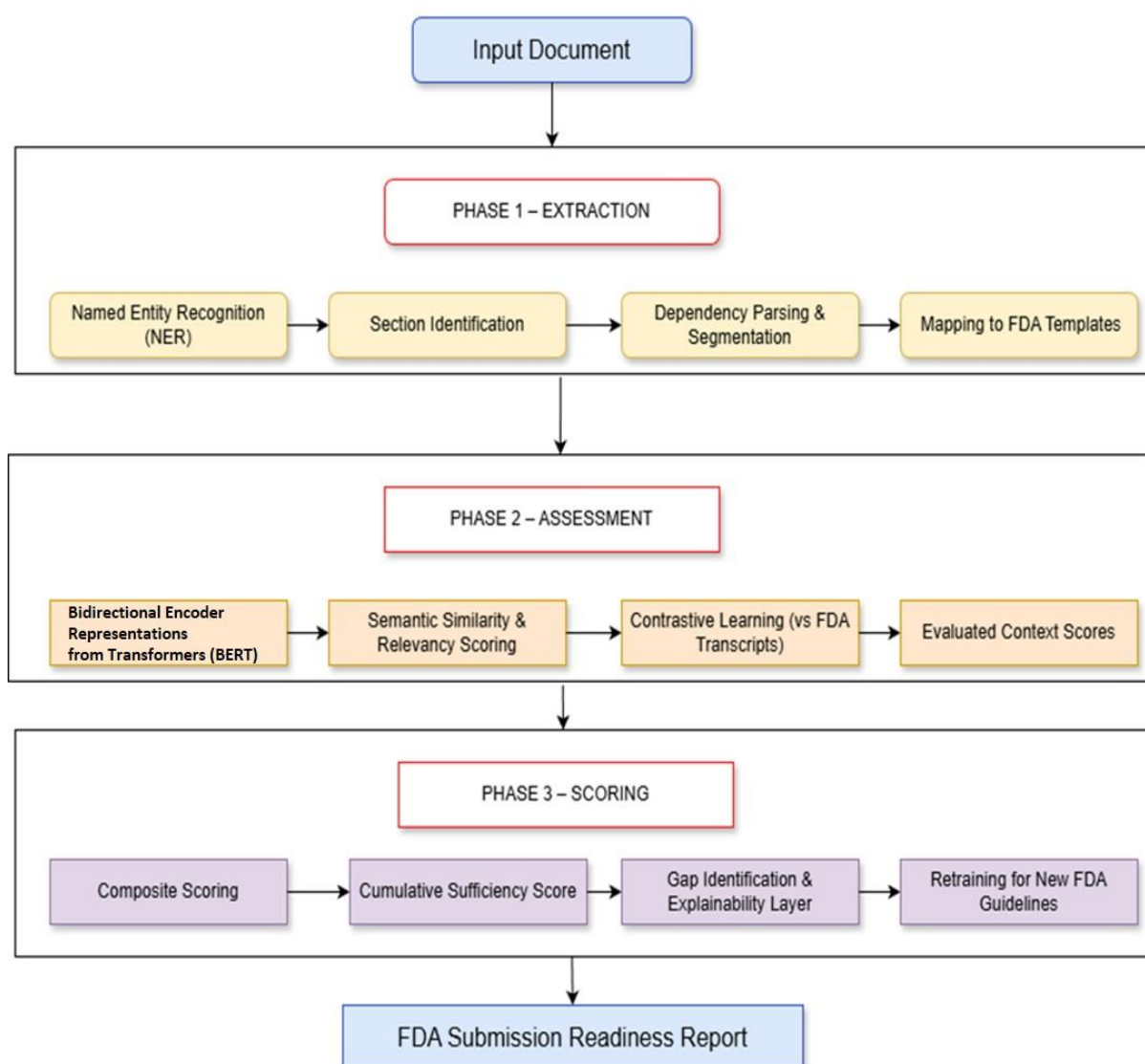




Figure 1: The proposed 3-phase NLP framework pipeline. The model details the Extraction (Phase 1), Assessment (Phase 2), and Scoring (Phase 3) stages used to generate an FDA Submission Readiness Report.

4. Integration of the NLP Framework into Regulatory Workflows

To ensure the usefulness of the application value of an NLP structure to determine the adequacy of data in submitting a meeting request to the FDA, it is necessary to have a seamless integration with the existing regulatory authoring and submission procedures. This type of system is most effectively used as part of the sponsor's Regulatory Information Management (RIM) ecosystem or document management systems (DMS), in which it could be a pre-submission assessment system and a post-submission analytics system.

Integration typically begins during the document authoring phase. The NLP framework can be implemented as a plug-in or deployed as a cloud-based microservice to perform real-time scanning of meeting request materials as regulatory teams prepare them. It can also integrate with content management platforms such as Veeva Vault, MasterControl, or SharePoint-based DMS environments to enable content analysis without disrupting the authoring interface. The real-time feedback loop enables regulatory writers to receive automated insights on missing information, inadequately justified questions, or areas of non-compliance with FDA expectations [20][21]. The NLP system can also conduct an extensive evaluation and generate a sufficient report at the moment the document is submitted to the pre-submission quality control (QC) section. This report is made up of document element breakdown, risk rates, the level of completeness, and provoking flags, which signify problematic language or unverified questions. To take an example, when the sponsor poses questions with regard to the practicability of a dosing schedule and fails to give related toxicological data, the NLP system is able to pinpoint the discrepancy and offer suggestions to make the submission harmonious, citing successful patterns of prior submissions [22]. The NLP system can also facilitate post-submission learning and improvement by analyzing FDA feedback and meeting minutes. By comparing the justifications and questions included in the submitted request with the regulator's responses, insights can be gained into which approaches lead to more constructive discussions or outcomes. This closed-loop learning mechanism enables the NLP framework to continuously refine its language models and scoring algorithms, progressively aligning with evolving regulatory expectations [23]. Also, integration extends to the level of generating dashboards for the regulatory affairs leadership. Examples of such dashboards can include portfolio-wide analytics, including the trends in submission readiness scores, frequent gaps, and performance patterns across therapeutic areas or the types of documents. This intelligence can also inform training programs, enabling updates to document templates and the refinement of future strategies for regulatory engagement [24]. The benefit of such a combined solution is that regulatory intelligence is operationalized, and it becomes embedded in the submission lifecycle. It reduced the level of individual reviewer expertise; they are scalable, and all requests for them are assessed on the adequacy of data that is expected to be provided to the FDA. After discussing the aspects of implementation, the discussion that follows is going to



discuss the validation and practical challenges in the implementation of such NLP setups in controlled environments.

5. Validation and Deployment Challenges of NLP in Regulatory Submissions

The usefulness of an NLP-based framework in the regulations working processes can be considered immense, but the implementation of this system is not without difficulties. The strong requirements to make such systems work and be in compliance with the regulatory standards require a rigorous validation process. Under a regulated environment, especially software systems that affect the content of a submission should be able to satisfy the validation requirements of the CFR, which provides assurance of data integrity, auditability, and reproducibility [25]. NLP model validation encompasses several dimensions. The first is performance validation, which involves testing the models on annotated datasets of FDA meeting requests and measuring precision, recall, and F1 scores (F1 score is a metric used in classification tasks to evaluate a model's accuracy by combining precision and recall into a single value) across content extraction, classification, and scoring tasks. This process requires high-quality labeled data, ideally including both approved and rejected FDA submissions, which are typically proprietary and difficult to obtain. Lack of data is a significant challenge, especially when it comes to educating the models to make decisions with high accuracy [26].

Secondly, the problem of explainability and transparency exists. The regulatory teams need to be in a position to interpret and have confidence in the outputs of the NLP system. Deep neural networks are not as acceptable as black-box models in a regulated environment unless they are complemented with explainable AI methods. Thus, the framework should contain the generation of reasons or attention visualization tools that explain why a section received a flag or a specific score. In the absence of this transparency, the adoption by regulatory professionals will be restricted [27]. Content variability is another significant threat. There is a wide range of FDA meeting requests depending on the type of therapeutic area, the type of submission, or the level of sponsor maturity. The one-size-fits-all model is not likely to be consistent across this range. The performance also has to be maintained, which results in the need for customization, fine-tuning, and domain adaptation, which consequently demand the professional work of AI and regulatory skills. Furthermore, new guidelines are being published or templates are getting updated, meaning that models are required to be retrained or updated, which requires the practice of a solid model lifecycle management [28].

Barriers to deployment are also security and confidentiality. The meeting request document has confidential information about the product and development. It is also of concern that the availability of NLP models in clouds exposes the data to breaches or the adherence to data protection laws like the General Data Protection Regulation (GDPR). On-premise deployment provides a different alternative at a higher cost and complexity of IT. This should be done through encryption, user access control, and audit trails so that the proprietary information is handled securely in any way [29]. Lastly, there is the human factor- resistance to the use of AI. The regulators, who are usually risk-averse, might find it difficult to base their judgment on algorithmic evaluations. This is where change management, training, and phase



implementation strategies come in, where users can develop trust in the system over time. It is possible to pilot the framework in less risky situations, like reviewing internal documents or a mock submission, to show value and not be at risk of regulation [30]. The struggles do not mean the removal of the value of the proposed NLP framework, but instead depict the need to consider the design, conduct a thorough validation, and implement the system to benefit the user. Taking these into account, the following section gives a description of future directions and opportunities of regulatory NLP systems in the larger drug development settings.

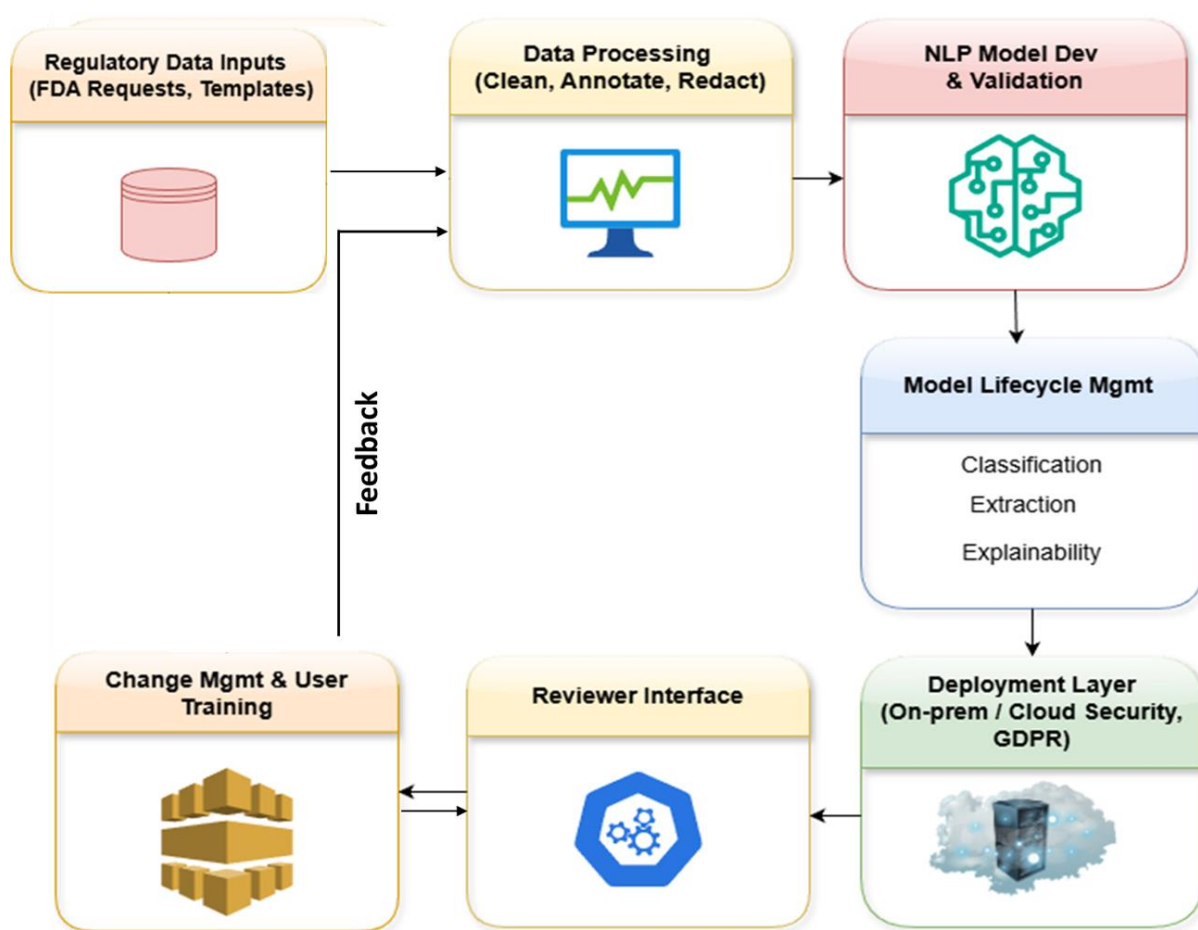


Figure 2: The figure illustrates a streamlined workflow for deploying an NLP system in a regulatory setting. It starts with regulatory data inputs (e.g., FDA templates), followed by data processing (cleaning, annotation, redaction), and model development. The process includes Model Lifecycle Management (Mgmt = Management) for classification, extraction, and explainability, leading to deployment in secure environments (on-prem/cloud, GDPR-



compliant). A reviewer interface supports user interaction, with outputs feeding into Change Management and training. A feedback loop ensures continuous system improvement.

6. Future Directions and Strategic Opportunities

The fact that NLP has been used in establishing the degree of data that is adequate to submit a request to the FDA to organize a meeting is not the sole element of a more profound shift in the functioning of the regulation. As the AI technologies continue to evolve, they will be used in other parts of the regulatory lifecycle, such as submission authoring and post-marketing surveillance. The further iterations of the proposed framework may involve predictive analytics, which may forecast the likelihood of the response by the FDA based on the patterns of the language, previous outcomes, and details specific to the area of therapeutic intervention. The further action that may be suggested at once is to enhance the functions of the framework to offer an assessment of complete briefing packages, and not just the initial meeting request letters. This would entail an advanced level of semantic analysis to correspond to supporting data tables, study summaries, and appendices to narrative justifications. Medical Dictionary for Regulatory Activities (MedDRA) or Clinical Data Interchange Standards Consortium (CDISC) can be considered to be domain-specific ontologies that can be incorporated into the system to improve standardization and uniformity in submissions. NLP systems may be combined with structured data repositories that include ClinicalTrials.gov, PubMed, and internal data lakes to counter claims when answering requests in a meeting with known datasets or literature, which will also enhance the credibility of documents. In this regard, the framework would not only turn into a reviewer but also a data verifier.

The other possible path is the use of conversational AI agents, which were trained on the history of FDA interactions. These agents can assist the sponsors in writing down questions that can raise regulatory concerns or refining clinical development plans in real-time. This, coupled with the NLP framework, is a move in the correct direction of smart co-authoring systems for regulatory professionals. The collaboration between the agencies, regulatory technology vendors, and sponsors will be instrumental in establishing the AI validation model that is specific to the NLP systems in regulations. This would use similar programs, like the Emerging Technology Program provided by the FDA or the Innovation Task Force of EMA, as a means of trying out and demonstrating such systems under a controlled environment before they become formally accepted and approved. Since the regulatory environment continues to be digitally transformed, the implementation of the NLP models like the one shown here offers a scalable, viable, and intelligent solution to an ancient problem in drug development, i.e., rendering the enterprise's critical regulatory filings readable, comprehensive, and persuasive. The pathway to be followed is that of measured innovation, which is grounded in science and tested by evidence and controlled by rigor.

Table of Abbreviations

Abbreviation	Full Form
--------------	-----------



AI	Artificial Intelligence
BERT	Bidirectional Encoder Representations from Transformers
CDISC	Clinical Data Interchange Standards Consortium
CFR	Code of Federal Regulations
DMS	Document Management System
FDA	Food and Drug Administration
GDPR	General Data Protection Regulation
IND	Investigational New Drug
Pre-IND	Pre-Investigational New Drug
MedDRA	Medical Dictionary for Regulatory Activities
NLP	Natural Language Processing
NER	Named Entity Recognition
QC	Quality Control
RIM	Regulatory Information Management
EMA	European Medicines Agency

References

1. Mak, K. K., & Pichika, M. R. (2019). Artificial intelligence in drug development: present status and future prospects. *Drug discovery today*, 24(3), 773-780.
2. Venna, S. R. (2023). AI and Automation in Regulatory Operations: The Future of eCTD Submissions. *Indo-American Journal of Pharma and Bio Sciences*, 21(2), 33-42.
3. Topol, E. J. (2019). High-performance medicine: the convergence of human and artificial intelligence. *Nature Medicine*, 25(1), 44-56.
4. Tiwari, P. C., Pal, R., Chaudhary, M. J., & Nath, R. (2023). Artificial intelligence revolutionizing drug development: Exploring opportunities and challenges. *Drug Development Research*, 84(8), 1652-1663.
5. Yang, Q., Liu, Y., Chen, T., & Tong, Y. (2019). Federated machine learning: Concept and applications. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 10(2), 1-19.
6. Oualikene-Gonin, W., Jaulent, M. C., Thierry, J. P., Oliveira-Martins, S., Belgodère, L., Maison, P., ... & Scientific Advisory Board of ANSM. (2024). Artificial intelligence



integration in the drug lifecycle and in regulatory science: policy implications, challenges and opportunities. *Frontiers in pharmacology*, 15, 1437167.

7. Vu, H. H., & Pariser, A. R. (2015). Pre-Investigational New Drug Meetings With the FDA: Evaluation of Meeting Content and Characteristics of Applications for New Drug and Biologic Products. *Therapeutic Innovation & Regulatory Science*, 49(1), 50–57. <https://doi.org/10.1177/2168479014558275>
8. GUDE, S., & GUDE, Y. S. (2024). The synergy of artificial intelligence and machine learning in revolutionizing pharmaceutical regulatory affairs. *Translational and Regulatory Sciences*, 6(2), 37-45.
9. Madan, S., Lentzen, M., Brandt, J., Rueckert, D., Hofmann-Apitius, M., & Fröhlich, H. (2024). Transformer models in biomedicine. *BMC medical informatics and decision making*, 24(1), 214.
10. Demner-Fushman, D., Chapman, W. W., & McDonald, C. J. (2009). What can natural language processing do for clinical decision support?. *Journal of biomedical informatics*, 42(5), 760-772.
11. Jin, Q., Yuan, Z., Xiong, G., Yu, Q., Ying, H., Tan, C., ... & Yu, S. (2022). Biomedical question answering: a survey of approaches and challenges. *ACM Computing Surveys (CSUR)*, 55(2), 1-36.
12. Lee, R. Y., Kross, E. K., Torrence, J., Li, K. S., Sibley, J., Cohen, T., ... & Curtis, J. R. (2023). Assessment of natural language processing of electronic health records to measure goals-of-care discussions as a clinical trial outcome. *JAMA Network Open*, 6(3), e231204-e231204.
13. Turchin, A., Masharsky, S., & Zitnik, M. (2023). Comparison of BERT implementations for natural language processing of narrative medical documents. *Informatics in Medicine Unlocked*, 36, 101139.
14. Oluwagbade, E. (2025). AI-Powered Regulatory Compliance: Automating Risk Assessment in Drug Approval Processes. *ResearchGate*.
15. Mayfield, E., & Rosé, C. P. (2020). *Biomedical Named Entity Recognition at Scale*. In A. Barros, A. Cardoso, & R. Monteiro (Eds.), *Advances in Artificial Intelligence – IBERAMIA 2020* (pp. 589–600). Springer. https://doi.org/10.1007/978-3-030-68763-2_48
16. Neumann, E. K., & Quan, D. (2006). Biodash: A semantic web dashboard for drug development. In *Biocomputing 2006* (pp. 176-187).
17. Singh, S. (2024). Leveraging AI and Machine Learning in Six-Sigma Documentation for Pharmaceutical Quality Assurance. *Chinese Journal of Applied Physiology*, 40, e20240005.



18. Han, Y., Ceross, A., & Bergmann, J. H. (2023). Uncovering regulatory affairs complexity in medical products: a qualitative assessment utilizing open coding and natural language processing (NLP). *arXiv preprint arXiv:2401.02975*.
19. Hsu, J. C., Wu, M., Kim, C., Vora, B., Lien, Y. T., Jindal, A., ... & Wu, B. (2024). Applications of advanced natural language processing for clinical pharmacology. *Clinical Pharmacology & Therapeutics*, 115(4), 786-794.
20. Higgins, D. C., & Johner, C. (2023). Validation of artificial intelligence containing products across the regulated healthcare industries. *Therapeutic Innovation & Regulatory Science*, 57(4), 797-809.
21. Bhatnagar, R., Sardar, S., Beheshti, M., & Podichetty, J. T. (2022). How can natural language processing help model-informed drug development?: a review. *JAMIA open*, 5(2), ooac043.
22. Ajmal, C.S., Yerram, S., Abishek, V. et al. Innovative Approaches in Regulatory Affairs: Leveraging Artificial Intelligence and Machine Learning for Efficient Compliance and Decision-Making. *AAPS J* 27, 22 (2025). <https://doi.org/10.1208/s12248-024-01006-5>
23. Parikh, R. B., Obermeyer, Z., & Navathe, A. S. (2019). Regulation of predictive analytics in medicine. *Science*, 363(6429), 810-812.
24. Ambalavanan, R., Snead, R. S., Marczika, J., Towett, G., Malioukis, A., & Mbogori-Kairichi, M. (2025). Ontologies as the semantic bridge between artificial intelligence and healthcare. *Frontiers in Digital Health*, 7, 1668385.
25. Zhou, K., Gattinger, G. The Evolving Regulatory Paradigm of AI in MedTech: A Review of Perspectives and Where We Are Today. *Ther Innov Regul Sci* 58, 456–464 (2024). <https://doi.org/10.1007/s43441-024-00628-3>
26. Chen, N., Lin, X., Jiang, H., & An, Y. (2024). Automated building information modeling compliance check through a large language model combined with deep learning and ontology. *Buildings*, 14(7), 1983.
27. Laparra, E., Mascio, A., Velupillai, S., & Miller, T. (2021). A review of recent work in transfer learning and domain adaptation for natural language processing of electronic health records. *Yearbook of medical informatics*, 30(01), 239-244
28. Rehan, H. (2023). AI-Powered Genomic Analysis in the Cloud: Enhancing Precision Medicine and Ensuring Data Security in Biomedical Research. *Journal of deep learning in genomic data analysis*, 3(1), 37-71.
29. Klaus, S., Van Hecke, R., Djafari Naini, K., Altingovde, I. S., Bernabé-Moreno, J., & Herrera-Viedma, E. (2022, July). Summarizing legal regulatory documents using



transformers. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 2426-2430).

30. Ajeigbe, K., & Owen, A. (2020). User-Centric Design of Cloud-Based Healthcare AI Applications.